Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-26 (2016) 147–152

# Gene Regulatory Network Inference Using Time-Stamped Cross-Sectional Single Cell Expression Data

**Nan Papili Gao*,** S. M. Minhaz Ud-Dean*,**
Rudiyanto Gunawan*,**,[1]**

*Institute for Chemical and Bioengineering, ETH Zurich, Zurich, Switzerland

** Swiss Institute of Bioinformatics, Lausanne, Switzerland

Abstract: In this paper we presented a novel method for inferring gene regulatory network (GRN) from time-stamped cross-sectional single cell data. Our strategy, called SNIFS (Sparse Network Inference For Single cell data) seeks to recover the causal relationships among genes by analyzing the evolution of the distribution of gene expression levels over time, more specifically using Kolmogorov-Smirnov (KS) distance. In the proposed method, we formulated the GRN inference as a linear regression problem, where we used Lasso regularization to obtain the optimal sparse solution. We tested SNIFS using *in silico* single cell data from 10- and 20-gene GRNs, and compared the performance of our method with Time Series Network Inference (TSNI), GEne Network Inference with Ensemble of trees (GENIE3), and an extension of GENIE3 for time series data called JUMP3. The results showed that SNIFS outperformed existing algorithms based on the Area Under the Receiver Operating Characteristic (AUROC) and Area Under the Precision-Recall (AUPR) curves.



*Keywords:* network inference, single cell, gene expression, gene regulatory network

## 1. INTRODUCTION

What is the gene regulatory network underlying the stochastic dynamics in cell fate decision-making process? How can we control such a process to give a desired outcome? The answers to these questions have important implications in understanding and manipulating the cell differentiation process, and more broadly, in providing new appreciation for the role of cell-to-cell variability in the fields of biology and medicine (Sandberg 2013). Thanks to recent advances in high-throughput cell profiling technologies, such as RNA-sequencing and real-time PCR (polymerase chain reaction), and in microfluidic devices for single cell handling, researchers now have the ability to obtain the transcriptional expression profile of a large set of genes for single cells. For example, Fluidigm Biomark® allows the measurements of the expression levels of 96 genes in 96 single cells in a single assay (Pieprzyk and High 2009). This device and several other single cell assay platforms offer much promise in overcoming the limitations of population-averaged measurements. In particular, population averages lack the necessary information for understanding biological processes, such as cell-fate determination, where cell-to-cell variability and stochastic dynamics play important functional roles (Stegle et al. 2015).

Many of the existing technologies for quantifying single cell gene expression, such as the Fluidigm Biomark®, rely on cell lysate assay, and therefore would not allow for a longitudinal single cell study. The analysis of single cell data requires new computational approaches, as many of the existing strategies do not explicitly account for and therefore are unable to take advantage of information contained in the cell-to-cell variability. For this reason, a plethora of new computational approaches have been introduced recently for the analysis of cross-sectional single cell data (Woodhouse et al. 2016). In particular, several methods focused on the reconstruction of the GRN using single cell transcriptomics based on Boolean networks (Chen et al. 2014; Moignard et al. 2015), stochastic modelling (Teles et al. 2013), gene co-expression/correlation (Kouno et al. 2013; Moignard et al. 2013; Pina et al. 2015), and nonlinear ODE systems (Ocone et al. 2015). The applications of these methods however faced a few challenges due to, for example, the requirement of dense time course data and high computational complexity that scales exponentially with the size of the network.

In this work, we proposed a novel method called SNIFS (Sparse Network Inference For Single cell data) for efficient inference of GRN from time-stamped cross-sectional single cell expression data. SNIFS begins with the computation of changes in the transcriptional expression distribution over time for each gene, quantified using the Kolmogorov-

2405-8963 
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.12.117

Smirnov (KS) distribution distances between two subsequent time points. The GRN inference involves finding a sparse linear model that describes the KS distance of a gene in a given time step as a function of the KS distances of all other genes in the previous time step. We adopted the Lasso regularization to find the optimal sparse solution (Tibshirani 1996).

We tested the performance of SNIFS to predict random subnetworks of *E. coli* and yeast GRN from *in silico* cross-sectional single cell transcriptional expression data. We compared the performance of SNIFS to that of a network inference method for time series (population-averaged) expression data TSNI (Time Series Network Inference) (Bansal et al. 2006), to GENIE3 (GEne Network Inference with Ensemble of trees) that uses a tree-based ensemble regression method (Huynh-Thu et al. 2010), and to JUMP3, a hybrid network inference algorithm that combines decision-trees with dynamical modeling for time series data (Huynh-Thu and Sanguinetti 2015). We showed that by employing for the time evolution of the gene expression distributions, SNIFS significantly outperformed TSNI, GENIE3 and JUMP3.

## 2. METHODS

### *2.1 GRN inference from single cell data using SNIFS*

In this subsection, we describe the method SNIFS. The subroutines of SNIFS are implemented in MATLAB (version 8.6; MathWorks Inc., Natick, MA, 2000) and available upon request.

Fig. 1 illustrates the GRN inference procedure using SNIFS. The input data for SNIFS (see Fig. 1A) are single cell readings of gene expression data at uniformly-spaced time points, for example from single cell real-time PCR. Let $s$ be the number of samples or individual cells, $m$ be the number of genes, and $n$ be the number of measurement time points.

The cross-sectional dataset comprises $n$ data matrices $D_{s\times m}$, where $D_{i,j}$ is the transcriptional expression value of gene $j$ in the $i$-th cell. The first step in SNIFS is the calculation of distribution distances of gene expression levels between time points (see Fig. 1B). In particular, we used the Kolmogorov-Smirnov distance (Massey 1951) to measure the changes in the gene expression distribution over time, as follows:

$$KS_{j,\Delta_k} = \max_d \left( \left| \hat{F}_{j,t_{k+1}}(d) - \hat{F}_{j,t_k}(d) \right| \right) \quad (1)$$

where the distance $KS_{j,\Delta t_k}$ denotes the KS distance of gene $j$ in the time window $\Delta t_k$, $\hat{F}_{j,t_k}(d)$ denotes the cumulative distribution function of gene $j$ constructed using the single cell expression data of gene $j$ at time point $t_k$ ($k$ = 1, 2, …, $n$) and $d$ denotes the (random) gene expression variable. Presently, we consider the scenario where single cell expression data are taken at evenly spaced time points. When data are only available at non-uniform time intervals, one could use interpolation algorithms (e.g. spline fitting) to generate artificial KS distances with uniform time meshes.

The basic premise of the GRN inference in SNIFS is that the KS distance of a gene at a given time step $\Delta t_k$ is proportional to the KS distances of its regulators from the previous time step $\Delta t_{k\text{-}1}$, where the proportionality coefficients indicate the strength of the gene regulations. Following this ansatz, we formulated the GRN inference as a multivariate linear regression problem, as follows: (see also Fig. 1C).

$$KS_{j,\Delta t_k} = KS_{1,\Delta t_{k-1}}\alpha_{1,j} + KS_{2,\Delta t_{k-1}}\alpha_{2,j} + \ldots + KS_{m,\Delta t_{k-1}}\alpha_{m,j} \quad (2)$$

where the regression coefficient $\alpha_{i,j}$ describe the influence of gene $i$ on gene $j$. For each gene, we thus solve a linear regression problem of the form: $\mathbf{y}=\mathbf{X}\boldsymbol{\alpha}$, where $\mathbf{y}$ denotes the $n-2$ vector of KS distances corresponding to time steps from $\Delta t_2$ to $\Delta t_{n-1}$, and $\mathbf{X}$ denotes the $(n-2)\times m$ matrix of KS distances corresponding to time steps $\Delta t_1$ to $\Delta t_{n-2}$, for all genes. The regression coefficients contained in the $m\times 1$ vector $\boldsymbol{\alpha}$ is further assumed to be sparse, that is each gene has only a small number of regulators (Chen et al. 1999).

As indicated in Fig. 1D, we employed the 'least absolute shrinkage and selection operator' (Lasso) regularization to obtain a sparse solution for $\boldsymbol{\alpha}$. Specifically, for the inference problem $\mathbf{y}=\mathbf{X}\boldsymbol{\alpha}$, we performed a penalized least square optimization as follows:

$$\min_{\alpha} \|Y - X\alpha\|_2^2 + \lambda\|\alpha\|_1 \quad \text{such that } \alpha \geq 0 \quad (3)$$

where the scalar weight $\lambda$ is data dependent and requires parameter tuning (Tibshirani 1996). Here, we used the MATLAB version of GLMNET algorithm to compute the Lasso regularization path, providing the vectors $\boldsymbol{\alpha}$ for different $\lambda$ values. GLMNET utilizes a cyclical coordinate descent method (Friedman et al. 2010), where the optimization above is iteratively performed one parameter at a time. In order to find the optimal weight $\lambda$, we implemented a leave-one-out cross validation (LOOCV). Briefly, in each trial, we generated the Lasso regularization path using the dataset excluding one sample that was assigned as the test set. Subsequently, we computed the mean squared error in predicting the test sample using $\boldsymbol{\alpha}$ vectors in this regularization path as a function of $\lambda$. Finally, we chose the optimal $\lambda$ value corresponding to either the minimum average prediction error $\lambda_{MIN}$ or the minimum average prediction error plus a standard error $\lambda_{SE}$. The final $\boldsymbol{\alpha}^*$ vector was taken from the regularization path generated using the entire dataset at the selected optimal $\lambda$ parameter.

### *2.2 TSNI*

We compared SNIFS and the use of time series single cell expression data, to a GRN inference algorithm using population-averaged data, called TSNI (Time Series Network Identification) (Bansal et al. 2006). TSNI relies on a linear