GPGene: Inferring Gene Regulatory Networks with Stochastic Variational Gaussian Processes

Naiqi Li[1], Ercan E. Kuruoğlu[1], Yong Jiang[1, 2] and Shu-Tao Xia[2]

(1) Tsinghua-Berkeley Shenzhen Institute (2) Tsinghua University

Introduction

- Goal:
  - Study how genes regulate each other from DNA expression data.
- Motivation:
  - Reveals the secret of life.
  - Broad applications in experimental design, drug design, etc.
  - Advanced DNA sequencing provides high throughput expression data with low cost and high efficiency.
- Main contributions:
  - Propose GPGene, a method for inferring gene regulatory networks.
  - In static Gene Regulatory Network (GRN) inference, GPGene outperforms GENIE3, the winner in the DREAM4 challenge.
  - In dynamic GRN inference, GPGene can capture the dynamic change of the network, and be used to predict future expression levels.

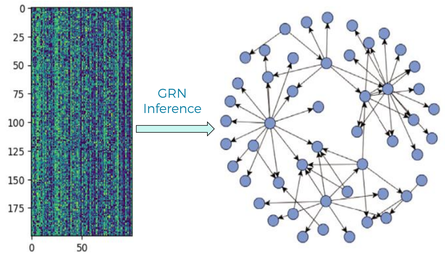

Figure 1. Given the observed gene expression data (left), the goal is to infer the underlying regulatory network (right).

Background

1. Stochastic Variational Gaussian Processes
   - Sparse GPs use a small set of inducing points to summarize all information. The joint PDF is factorized as:

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = \underbrace{p(\mathbf{f}[\mathbf{u}; \mathbf{X}, \mathbf{Z}) p(\mathbf{u}; \mathbf{Z})}_{\text{GP prior}} \underbrace{\prod_{i=1}^{N} p\left(y_i | f_i\right)}_{\text{likelihood}},$$

   - Inducing points are computed by introducing a distribution q to approximate the posterior, and optimizing the ELBO:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{f}, \mathbf{u})}\left[\log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{u})}{q(\mathbf{f}, \mathbf{u})}\right] = \sum_{i=1}^{n} \mathbb{E}_{q(f_i)}\left[\log p\left(y_i | f_i\right)\right] - \mathrm{KL}\left[q\left(\mathbf{u}\right) \| p\left(\mathbf{u}\right)\right].$$

   - The optimization can be achieved via sampling [1].
2. Automatic Relevance Detection Kernel
   - The Automatic Relevance Detection (ARD) kernel can automatically discover which features are important for the prediction.
   - Linear ARD kernel: $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{p} \sigma_i^2 x_i x_i'$.
   - Polynomial ARD kernel: $k^d(\mathbf{x}, \mathbf{x}') = (\sum_{i=1}^{p} \sigma_i^2 x_i x_i' + 1)^d$.
   - $k^d(\mathbf{x}, \mathbf{x}') = \left\langle \varphi^d(\mathbf{x}), \varphi^d(\mathbf{x}') \right\rangle$,
     $\varphi^2(\mathbf{x}) = (\sigma_1^2 x_1^2, ..., \sigma_p^2 x_p^2, \sqrt{2}\sigma_p \sigma_{p-1} x_p x_{p-1}, ..., \sqrt{2}\sigma_2 \sigma_1 x_2 x_1, \sqrt{2}\sigma_p x_p, ..., \sqrt{2}\sigma_1 x_1, 1)^\top$.
3. Related Work

Title	Year	Method	Nonlinear	Nonstationary
Time-Dependent Gene Network Modelling by Sequential Monte Carlo (S. Ancherbak et al.)	2016	Particle filtering	Yes	Yes
Tracking of time-varying genomic regulatory networks with a LASSO-Kalman smoother (J. Khan et al.)	2014	Kalman filtering + LASSO	No	Yes
Inferring gene regulatory networks with nonlinear models via exploiting sparsity (A. Noor et al.)	2012	Particle/Kalman filtering + LASSO	Yes	Yes
Estimating time-varying networks (M. Kolar et al.)	2010	Markov Random Fields	Partially	Yes
Inferring regulatory networks from expression data using tree-based methods (V. Huynh-Thu et al.)	2010	Regression	No	No
ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context (A. Margolin et al.)	2006	Information theory	Yes	No

Methodology

- GRN Inference as Feature Selection
  1. Suppose there are p genes. Expression level of the k-th subject is $\mathrm{x}^{(k)}$.
  2. Denote all the genes except for j as $\mathrm{x}_{-j}^{(k)} = \left\{x_1^{(k)}, ..., x_{j-1}^{(k)}, x_{j+1}^{(k)}, ..., x_p^{(k)}\right\}$.
  3. Use sparse GP to learn to predict the j-th gene: $f\left(\mathrm{x}_{-j}^{(k)}\right) = x_j^{(k)}$.
  4. ARD kernel automatically learn which genes are relevant.
  5. If gene i is very relevant in predicting j, $i \rightarrow j$ exist in the network.
- Flexible Feature Selection with ARD Kernels
  1. The most natural and intuitive choice is the linear kernel.
  2. Linear functions cannot explain all regulatory relations. See Figure 2.
  3. Polynomial kernel with degree 2 can model the situation in Figure 2.

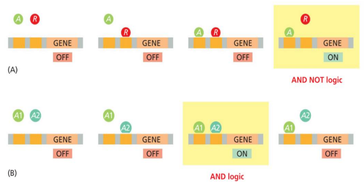

Figure 2. Target gene (GENE) is controlled by A (activator) and R (repressor). In (A) the target gene is activated if A is binding to the promoter and R is not. Figure is taken from [2].

- Dynamic GRN Inference with Time-dependent Sampling

Motivation: GRN is a time-varying.

- We introduce a time-varying weighted sampling scheme.
- For example when t=190, data within [150, 190] has greater importance in deciding the network structure.

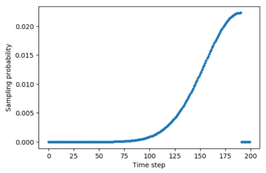

Figure 3. Time-dependent sampling distribution. Data closed to current time has higher probability to be sampled.

The theorem states that with a time dependent weighted likelihood, the corresponding ELBO can be obtained by time-dependent sampling:

Theorem. Let time-dependent likelihood be $\tilde{p}(y_i|f_i, t) = \frac{1}{C_i}\mathcal{N}(y_i|f_i, \sigma^2)^{w(t_i;t)}$, then the corresponding ELBO is:

$$\hat{\mathcal{L}}_t = \mathrm{E}_{w(t_i;t)}\left[\mathrm{E}_{q(f_i)}\left[\log \mathcal{N}(y_i|f_i, \sigma^2)\right]\right] - \mathrm{KL}[q(\mathbf{u})||p(\mathbf{u}; \mathbf{Z})] - K,$$

where K is a constant.

Dataset

- We use GeneNetWeaver [3, 4] to generate our synthetic dataset.
- We consider multifactorial data, which are the steady-state expression levels resulted from multifactorial perturbations in the initial state.
- Networks are extracted from the known GRNs of *E.coli* and *S.cerevisiae*.
- The expression data is generated by dynamic models based on ODE.

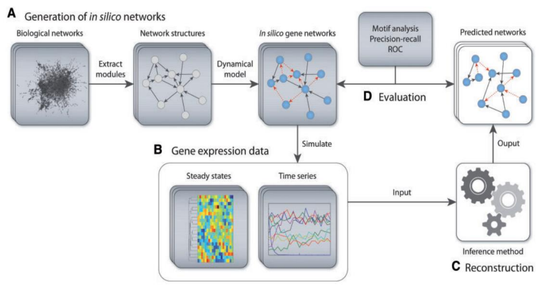

Figure 4. Data generation and evaluation procedure of GeneNetWeaver. Figure is taken from [4].

Experimental Results

1. Static GRN Inference

We first demonstrate the ability of GPGene in inferring static GRNs.

- We make comparison with GENIE3 [5], which is the winner of the DREAM4 challenge. Metrics (larger means better):
  - AUROC: Area under the ROC curve.
  - AUPR: Area under the precision-recall curve.

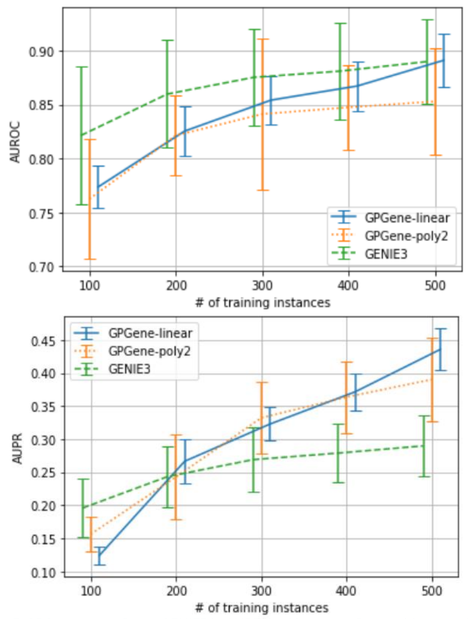

Figure 5. GPGene outperforms GENIE3 in AUPR, and improves faster with more data. GPGene with linear kernel outperforms the nonlinear one. We suspect that this is because GNW does not fully model the complex nonlinear interactions of multiple genes during the data synthesis,

2. Dynamic GRN Inference

As there is no widely accepted benchmarks, we experiment with toy data. Network is changed at t = 100.

- t = 1, ..., 100:
- t = 101, ..., 200:

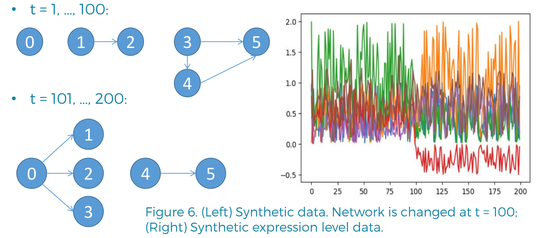

Figure 6. (Left) Synthetic data. Network is changed at t = 100; (Right) Synthetic expression level data.

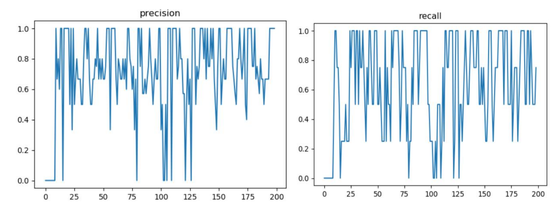

Figure 7. Precision and recall results change with time.

3. DNA Expression Prediction during Dynamic GRN Inference

GPGene naturally can be used to predict the gene expression levels in the future.

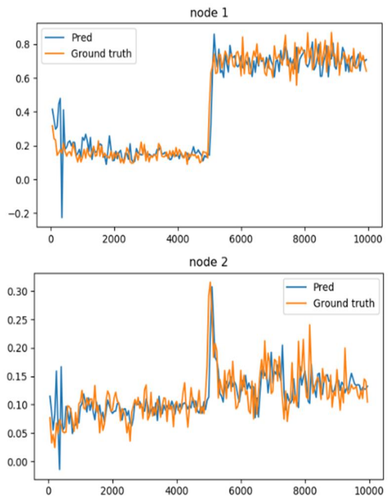

Figure 8. Predictive expression levels of two genes. An abrupt external perturbation is introduced at the middle of the experiment. Note that at the change point the prediction lags one step behind, but it can immediately adapt itself.

Conclusion

- We propose GPGene, a novel method for inferring gene regulatory networks (GRNs) with stochastic variational Gaussian processes (SVGPs).
- The key idea is to cast the GRN inference problem into a feature selection task during the regression.
- We use SVGPs to predict the gene expression levels during the regression, and utilize ARD kernel for feature selection.
- The rationality behind the linear and polynomial kernels was explained.
- GPGene can capture the dynamic change of the network structure by introducing a time-dependent weighted likelihood,.
- We show that the optimization of the ELBO can be efficiently achieved by a time-dependent weighted sampling scheme.
- Thorough experimental study was performed to demonstrate the effectiveness of our method.
- For static GRN inference our method outperforms GENIE3, which was the winner in the multifactorial track of the DREAM4 challenge.
- For the dynamic GRN inference problem, it is showed that our method can not only capture the dynamic change of the network, but also be used to predict future expression levels with satisfactory accuracy.

References

- [1] H. Salimbeni and M. Deisenroth, "Doubly stochastic variational inference for deep Gaussian processes," in Advances in Neural Information Processing Systems (NIPS-2017), 2017, pp. 4588–4599.
- [2] B. Alberts, J. Wilson, A. Johnson, T. Hunt, J. Lewis, K. Roberts, M. Raff, and P. Walter, Molecular Biology of the Cell. Garland Science, 2008.
- [3] R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky, "Towards a rigorous assessment of systems biology models: the DREAM3 challenges," PloS one, vol. 5, no. 2, p. e9202, 2010.
- [4] T. Schaffter, D. Marbach, and D. Floreano, "GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods," Bioinformatics, vol. 27, no. 16, pp. 2263–2270, 2011.
- [5] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," PLoS ONE, vol. 5, no. 9, p. e12776, 2010.