Mathematical Fundations on Gene Regulatory Network Inference

Supervised by Dr Shan He
Presented by Dong Li
2015-11-18

Outline

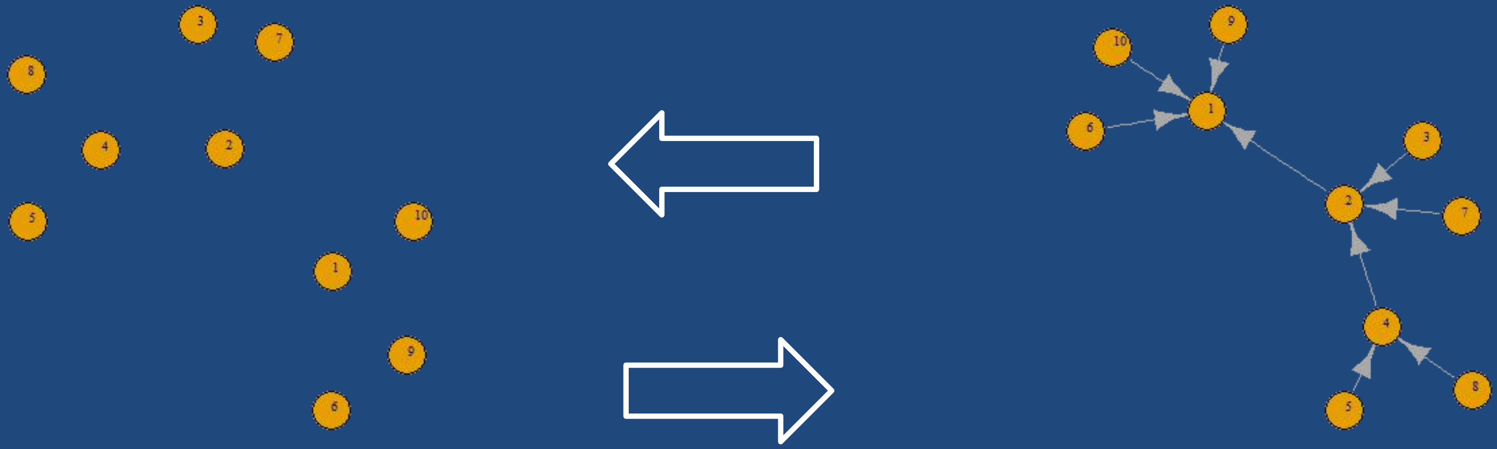
- Background
- Methods
  - Classic ones
  - GINIE3
  - Ensemble
- Case study: step-by-step

Background

- GRNs are simplified representations of regulatory mechanisms: G=(V,E)
- GRN inference is a fundamental task for systems biology
- Known as network identification, or reverse engineering
- Relationship with co-expression network

Inference is hard

- Goal: recovering GRN from expression data



genes

inferred network

Problem Definition

- Find regulatory relationships from expression data

$$gene_i \xrightarrow{w_{ij}} gene_j$$

- Gene expression matrix

$$X \in \Re^{n \times p}, n << p$$

- F: Mapping from X to G=(V,E,W)

Metrics

- For regression inference methods
  - least square error for gene i in all exprs

$$\sum_{j=1}^{N}\left(x_i^j - \sum_{k\neq i} x_k^j w_{k,i}^j - \varepsilon_i\right)^2$$

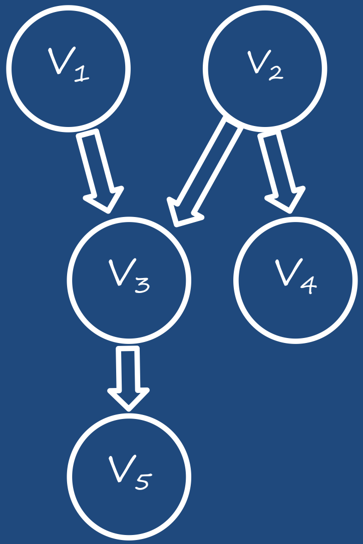
- For inference quality
  - AUPR, AUROC, self-defined scores...
- For biological intepretation

Methods

- Bayesian Networks
- Boolean Networks
- Differential Equations
- Regression Models

Bayesian Networks

- History
  - From cluster to structure
- Graphical probabilistic models
- Two-stage process of learning
  - model selection
  - parameter fitting (EM)
- Casual network = Bayesian network only when casual Markov assumption holds



$$P(V) = \prod_{i=1}^{p} P(V_i \mid Pa(V_i))$$

$$P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)} \propto P(D \mid G)P(G)$$

$$P(D \mid G) = \int P(D \mid G, \Theta)P(\Theta \mid G)d\Theta$$

$$\approx BDeu(G)$$

Boolean Networks

- Directed Graph where each node is boolean variable associated with a Boolean function e.g. $f_i(x_1, x_2) = x_1 \; or \; (not \; x_2)$
- Time (different conditions) as states to be transited

$$S(t) = (x_1(t), x_2(t), \ldots, x_n(t))$$

- Reverse engineering: given observation of states and want to get the network
- Probabilistic BN: more realistic, similar but different to Bayesian networks

Differential Equations

- DE: quantifying the rate of change

$$\frac{dx_i}{dt} = f_i(x_{i1}, x_{i2}, \ldots, x_{il})$$

- Linear assumption on function f

$$\frac{dx_i(t)}{dt} = f_i(x_{i1}, x_{i2}, \ldots, x_{il}) = \sum_{j=1}^{l} w_{ij} x_{ij}(t)$$

- differential approximated by

$$\frac{dx_i}{dt} \approx \frac{\Delta x_i}{\Delta t} = \frac{x_i(t+1) - x_i(t)}{\Delta t}$$

- Weight matrix = graph, solved by LR

$$\frac{d}{dt} X_{n \times p} = W_{n \times n} X_{n \times p} + E_{n \times p}$$

Regression Models

- Decompose it into p subproblems, each takes one gene as target
- Regression in experiment j for gene i

$$x_i^j = f_i(x_1^j, \ldots x_{i-1}^j, x_{i+1}^j, \ldots x_p^j) + \varepsilon_i$$

- Linear assumption on f

$$x_i^j = x_1^j w_{1,i}^j + \ldots + x_{i-1}^j w_{i-1,i}^j + x_{i+1}^j w_{i+1,i}^j \ldots + x_n^j w_{p,i}^j + \varepsilon_i$$

- Final weights from aggregation

Single gene regression

- Linear asumption with sparsity
  - Lasso selector

$$w = \arg\min_{w} \| Y - Xw \|^2 + \lambda \| w \|_1$$

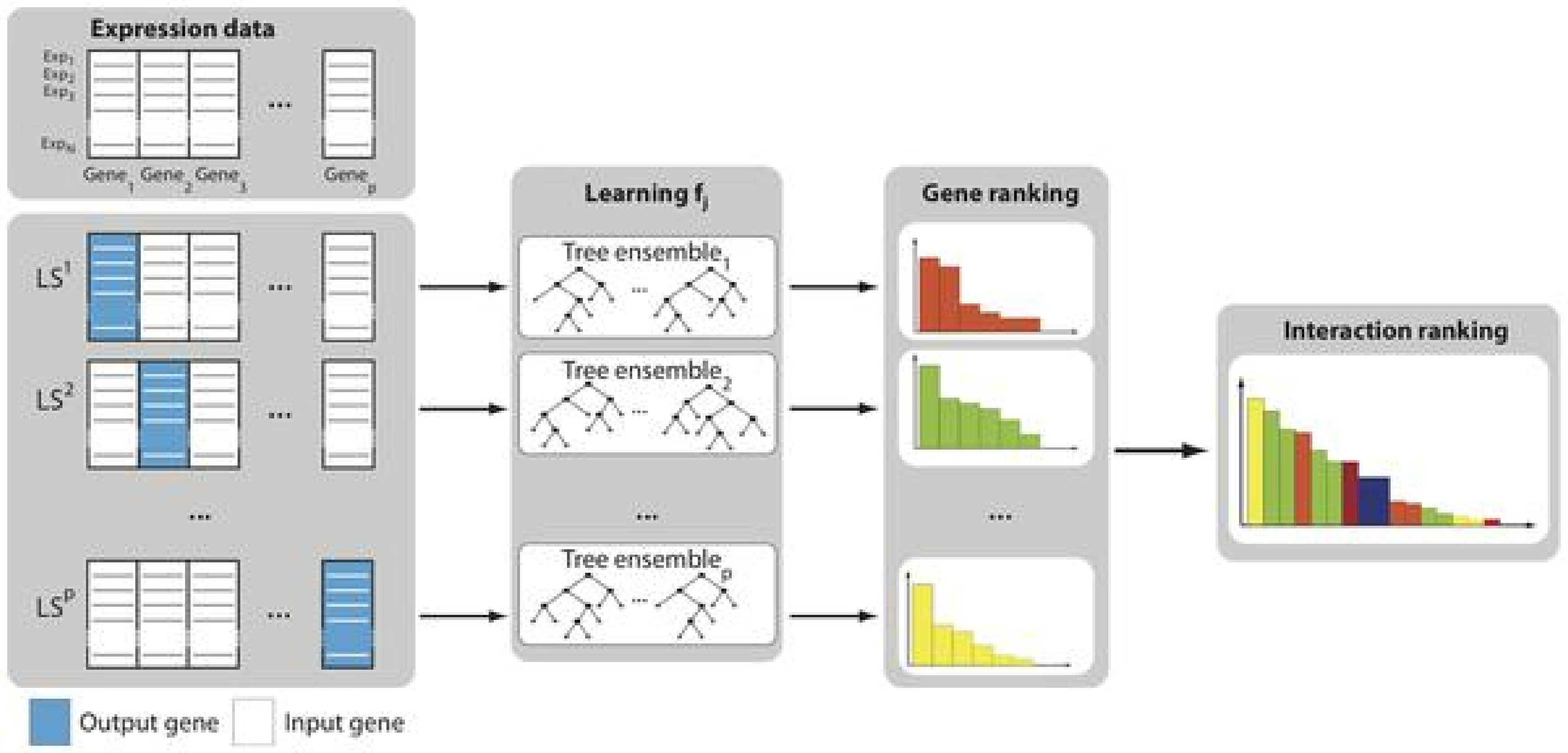
  - Dantzig selector

$$w = \arg\min_{w} \|w\|_1 \ s.t. \left\| X^T (Y - Xw) \right\|_{l_\infty} \leq \delta$$

- Non-linear, e.g. tree model
  - CART
  - nodes importance as weights

$$I(N) = \# S Var(S) - \# S_t Var(S_t) - \# S_f Var(S_f)$$

- *Figure 1. GENIE3 procedure.*



Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE 5(9): e12776. doi:10.1371/journal.pone.0012776
http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0012776

Beyond ensemble

- Meta analysis
  - combine different algorithms, e.g. regression+basyesian networks+...

    Vignes, Matthieu, et al. "Gene regulatory network reconstruction using Bayesian networks, the Dantzig Selector, the Lasso and their meta-analysis."*PloS one* 6.12 (2011): e29165.
- Wisdom of crowds
  - community based methods, combine all
  - selective based on diversity

    Marbach, Daniel, et al. "Wisdom of crowds for robust gene network inference. " *Nature methods* 9.8 (2012): 796-804.

Discussion

- Curse of dimensionality
  - $n << p$ makes methods unrealiable
  - Seems no hope to solve from this view

- No ground truth
  - If there is, we can measure the quality
  - If there is, what the inference task for

Case study with R

- Data preparation
- Network inference GENIE3
- Visualization using igraph package
- Biological intepretation (not available)

Thanks
Q/A?