Priors in Dependency network learning

Sushmita Roy
sroy@biostat.wisc.edu

Computational Network Biology
Biostatistics & Medical Informatics 826
https://compnetbiocourse.discovery.wisc.edu

Oct 11th 2018

Goals for this lecture

- Incorporating priors in Dependency networks using linear regression
- Incorporating priors in Dependency networks using tree models

Prior-based approaches for network inference

- Given
  - Gene expression data and
  - Complementary data that supports the presences of an edge
    - Presence of a sequence motif on a gene promoter
    - ChIP-chip/seq binding of factor X on gene Y's promoter
- Do
  - Predict which regulators drive the expression of a target gene, while incorporating complementary evidences as much possible
- How?
  - Place a prior on the graph where the prior is obtained from complementary data

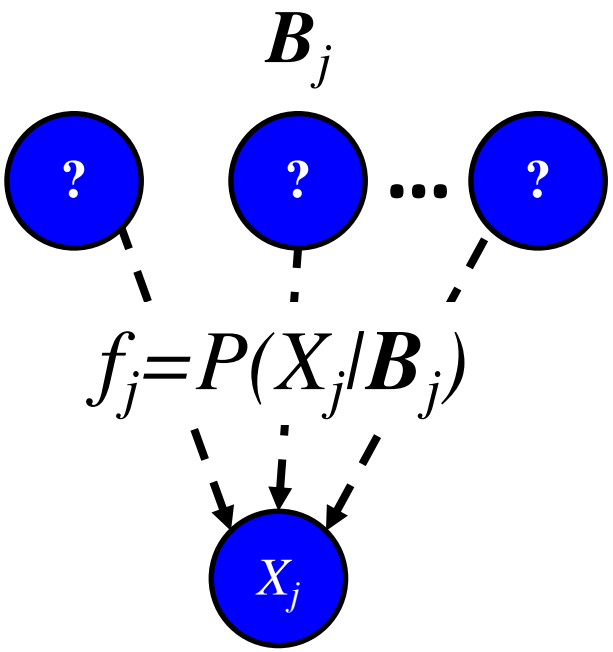
Recall Dependency network

- A type of probabilistic graphical model
- Approximate Markov networks
  - Are much easier to learn from data
- As in Bayesian networks has
  - A graph structure
  - Parameters capturing dependencies between a variable and its parents
- Unlike Bayesian network
  - Can have cyclic dependencies
  - Computing a joint probability is harder
    - It is approximated with a “pseudo” likelihood.

Dependency Networks for Inference, Collaborative Filtering and Data visualization
Heckerman, Chickering, Meek, Rounthwaite, Kadie 2000

Learning dependency networks

- One can think about this problem as estimating the Markov blanket of each random variable



- Let B_j denote the Markov Blanket of a variable X_j.
- B_j is the set of variables that make X_j independent of all other variables, X_{-j}

$$P(X_j | \mathbf{X}_{-j}) = P(X_j | \mathbf{B}_j)$$

- B_j can be estimated by finding the set of variables that best predict X_j
- This requires us to specify the form of $P(X_j | B_j)$

Overview of the Inferelator algorithm

- Based on linear regression models
- Handles time series and steady state data
- Prior is incorporated at the edge weight using two strategies
  - Modified Elastic Net
  - Bayesian Best Subset Regression

Greefield et al. 2013, Bonneau et al. 2007

Notation

- p regulators
- x_i: Gene expression levels for the i^{th} gene
  - $x_i(t)$: Expression of gene i at time t/sample t
- R: Number of samples
- β: Regression weight vector

Modeling the relationship between regulator and target in Inferalator

- Time series

Change in expression with time

$$y_i(t_{k+m}) = \sum_{p \in P_i} \beta_{i,p} x_p(t_k)$$

m is the time lag

$$i = 1, \ldots, N, \quad k = 1, \ldots, K-1$$

- Steady state

$$x_i(e_l) = \tau_i \sum_{p \in P_i} \beta_{i,p} x_p(e_l),$$

$$i = 1, \ldots, N, \quad l = 1, \ldots, L$$

Number of genes Number of samples

Network inference: Estimate coefficients $\beta_{i,p}$

Two approaches to integrate prior graph structure

- Modified Elastic Net (MEN)

- Bayesian Best Subset Regression (BBSR)

Recall regularized regression

- The regularized regression framework can be generally described as follows:

Regularization term

$$\text{minimize}_{\beta_0,\beta_i} \left[\frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right] + \lambda f(\beta)$$

Depending upon *f* we may have different types of regularized regression frameworks

Regularized regression

- $f(\beta)$ takes the form of some norm of β
- L1 norm

$$||\beta||_1 = \sum_{j=1}^{p} |\beta_i|$$

- L2 norm

$$||\beta||_2^2 = \sum_{j=1}^{p} \beta_i^2$$

- Elastic net

$$\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1$$

Elastic net regression

- If there are correlated predictors, LASSO will arbitrarily decide between the two to include or exclude
- Elastic net provides a tradeoff between ridge and LASSO.

Elastic net regression

- Elastic net regression objective for the i^{th} gene

$$\text{minimize}_{\beta} \frac{1}{2} \sum_{r=1}^{R} \left(y(r) - \sum_{p} \beta_p x_p(r) \right)^2 + \lambda \left[\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \right]$$

- Which can be equivalently written as

Minimize

$$\sum_{r=1}^{R} \left| y_i(r) - \sum_{p \in P_i} \beta_{i,p} x_p(r) \right|^2$$

Subject to

L1 norm L2 norm

$$(1-\xi) \sum_{p \in P_i} |\beta_{i,p}| + \xi \sum_{p \in P_i} \beta_{i,p}^2 \leq s_i \sum_{p \in P_i} |\beta_{i,p}^{\text{ols}}|$$

Estimate via cross validation

Modified Elastic Net (MEN)

- The modification to Elastic net

$$(1-\xi)\sum_{p\in P_i}|\theta_{i,p}\beta_{i,p}|+\xi\sum_{p\in P_i}\beta_{i,p}^2 \le s_i\sum_{p\in P_i}|\beta_{i,p}^{\text{ols}}|$$

Set this <1 so that if there is a prior edge between x_p->y_i, the regression coefficient will be penalized less

Two approaches to integrate prior graph structure

- Modified Elastic Net (MEN)

- Bayesian Best Subset Regression (BBSR)

Probabilistic interpretation for the one predictor case

- Recall our linear model for one predictor

$$y_i = x_i\beta_1 + \epsilon_i$$

Noise

- Assume noise is distributed according a Gaussian with mean 0 and variance σ

$$y_i \sim \mathcal{N}(x_i\beta_1, \sigma)$$

- How to estimate β_1 from N datapoints?
  - Maximize likelihood of data given model

Maximum Likelihood estimate of β_1

- Likelihood of data

$$LL = \prod_{i=1}^{N} P(y_i | x_i, \beta_1, \sigma)$$

$$LL = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - x_i\beta_1)^2}{2\sigma^2}\right)$$

Taking log

$$= \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} - \sum_{i=1}^{N} \frac{(y_i - x_i\beta_1)^2}{2\sigma^2}$$

Deriving wrt β_1 and setting to 0

$$\beta_1 = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} {x_i}^2}$$

Would get the same answer if minimizing RSS

Probabilistic interpretation in case of p inputs

- Assume output Y is

$$y_i \sim \mathcal{N}(\mathbf{x}_i\beta, \sigma)$$

- Again can compute likelihood, maximize it to find β
- Again the ML estimate would be the same as we derived by minimizing the RSS

Bayesian framework to estimate parameters

- Instead of optimizing the likelihood, we put a prior on the parameters and optimize the posterior probability of the parameters

$$P(\beta|\mathbf{D}) \propto P(\mathbf{D}|\beta)P(\beta)$$

Gaussian data likelihood

Parameter prior

What types of priors can we use?

Priors on parameters in regression

- Gaussian prior

$$P(\beta) = \mathcal{N}(0, \tau^2 \mathbf{I})$$

$$P(\beta) \propto \exp(-\frac{\beta^T \beta}{2\tau^2})$$

  - Also called ridge regression

- Laplace prior $P(\beta_i) = \text{Laplace}(0, t)$

$$P(\beta_i) \propto \exp(-\frac{|\beta_i|}{t})$$

  - Also called Lasso regression

Bayesian Best Subset Regression (BBSR)

- Based on a Bayesian framework of model selection
  - Search among all subsets of regulators and pick the best one to minimize trade off between data fit and model complexity
- Assume that the expression level y is distributed according to a Gaussian distribution

$$\left(y|\beta,\sigma^2,X\right) \propto \mathrm{N}_n\left(X\beta,\sigma^2\mathrm{I}\right)$$

Response variable

Regulators

Prior over parameters is a Gaussian

- Place a prior distribution on parameters, and incorporate prior knowledge of interactions in the parameters

$$p(\beta|\sigma^2) \propto N_n(\beta^0, g(X'X)^{-1}\sigma^2)$$

Prior

A number between 0 and infinity $g \in (0,\infty)$

BBSR continued

- The posterior distribution over the parameters is given as:

$$p(\beta|y,\sigma^2) \propto \mathcal{N}(\frac{1}{g+1}\beta^0 + \frac{g}{g+1}\beta^{OLS}, \sigma^2\frac{g}{g+1}(X'X)^{-1})$$

- g can be tuned to provide a trade-off between the prior and the OLS solution
- When g is larger, beta is closer to the OLS solution
- When it is smaller, beta is closer to the prior
- The prior is set to be a vector of all 0s

BBSR continued

- Inferelator uses a p-dimensional vector for p predictors

$$\bar{g} = \{g, \cdots, \frac{1}{g}, g, \cdots, \frac{1}{g}\}, \text{where } g > 1$$

Predictors with prior are set to g (push more towards the OLS solution)

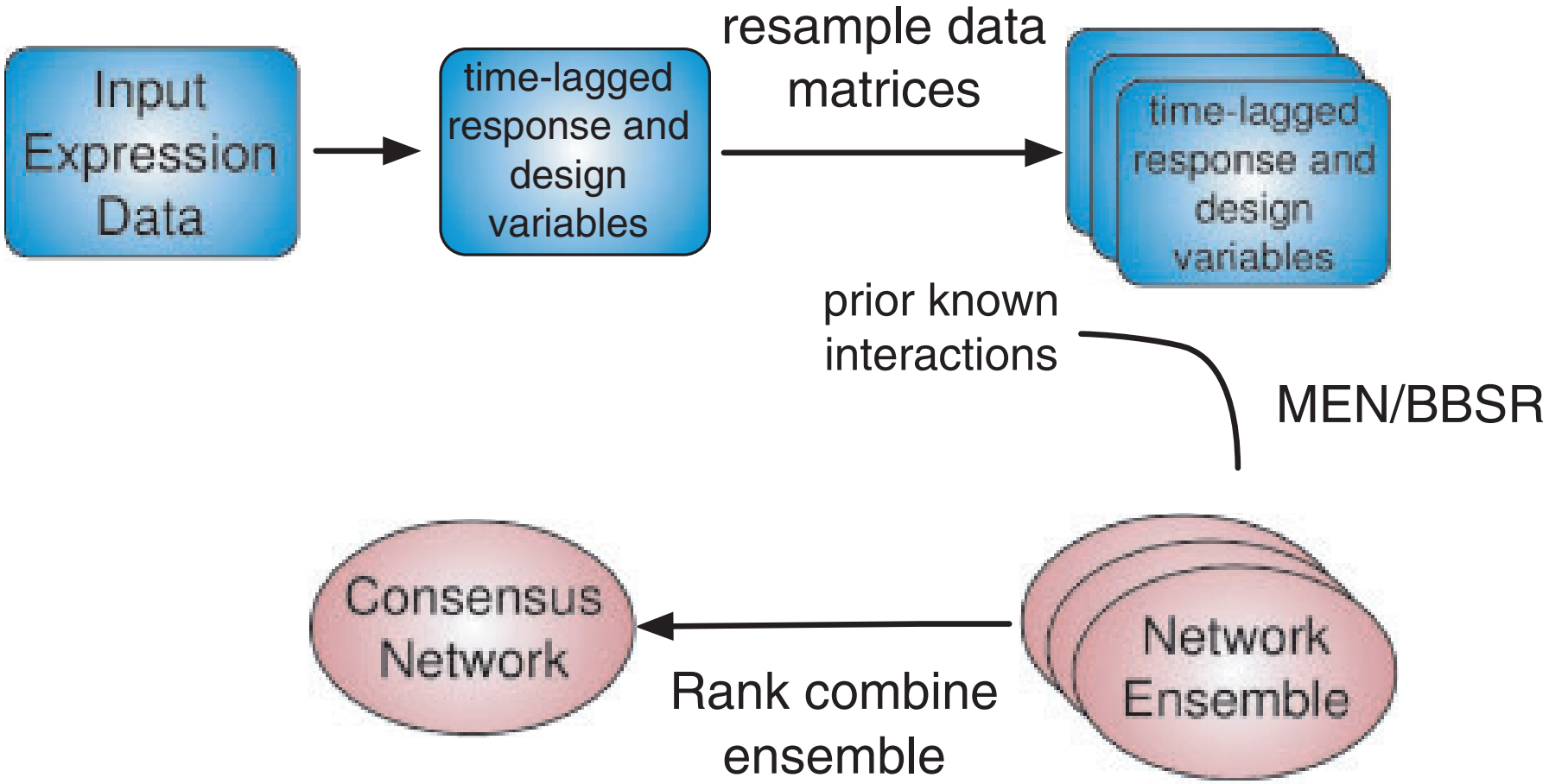
BBSR model selection

- The final step in BBSR is to determine the best model out 2^p possible sets
- p cannot be very high: the approach sets p to 10
- The best model is the one that minimizes prediction error and has the lowest model complexity

Experimental setup

- Three datasets
  - DREAM4: In silico dataset with 100 nodes
  - *E. coli* dataset from DREAM5
  - *B. subtilis* dataset
- Evaluation based on AUPR
  - Ranking of edges obtained from a bootstrapping strategy
- Questions asked
  - How does the prior parameter affect the performance?
  - Does the prior hamper performance on parts of the network without prior support?
  - How robust is the framework to noisy priors?

Workflow of experiments

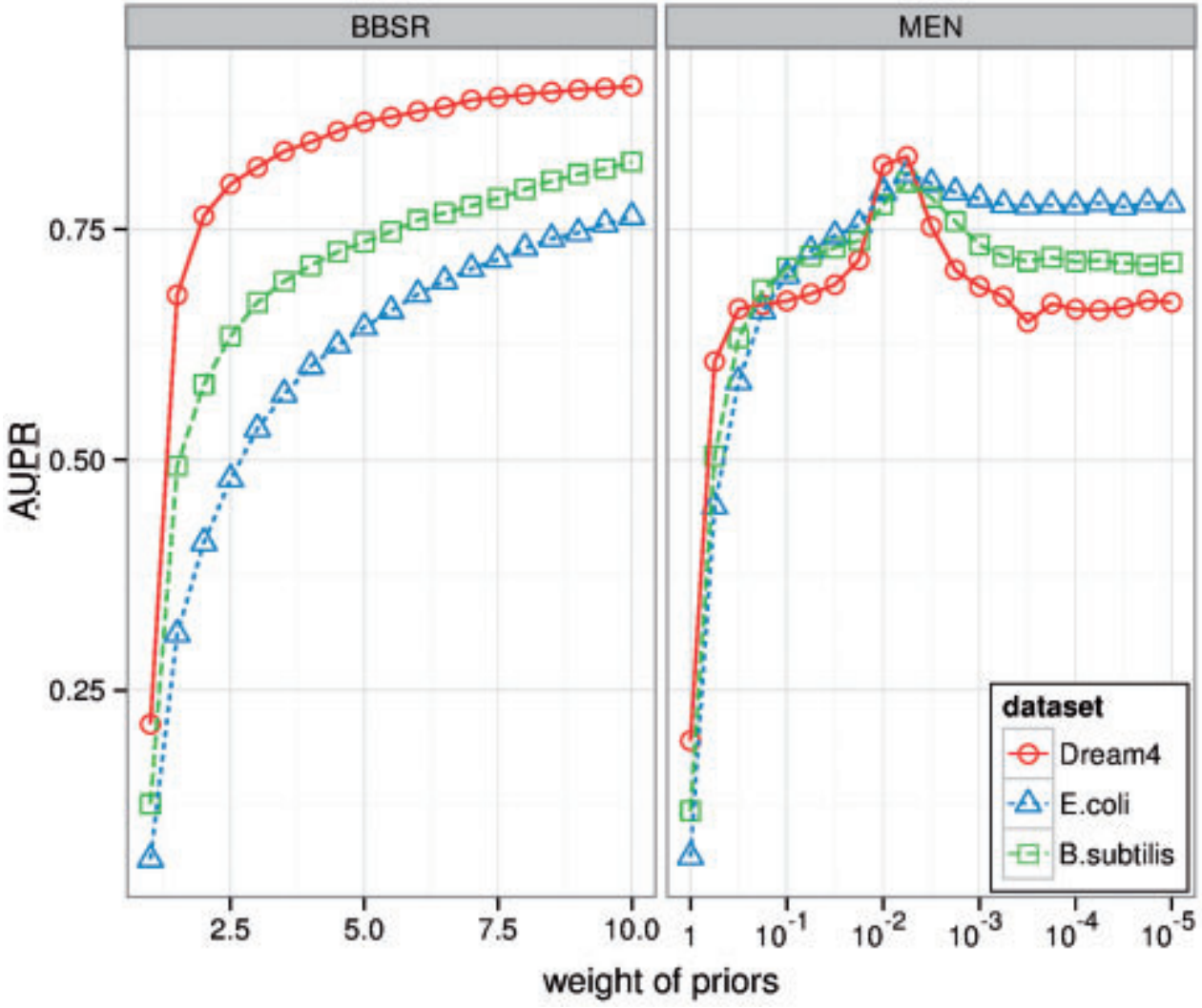


Key questions asked in experiments

- How does one pick the prior parameter values?

- How does one identify novel edges and not incorporate only the prior?

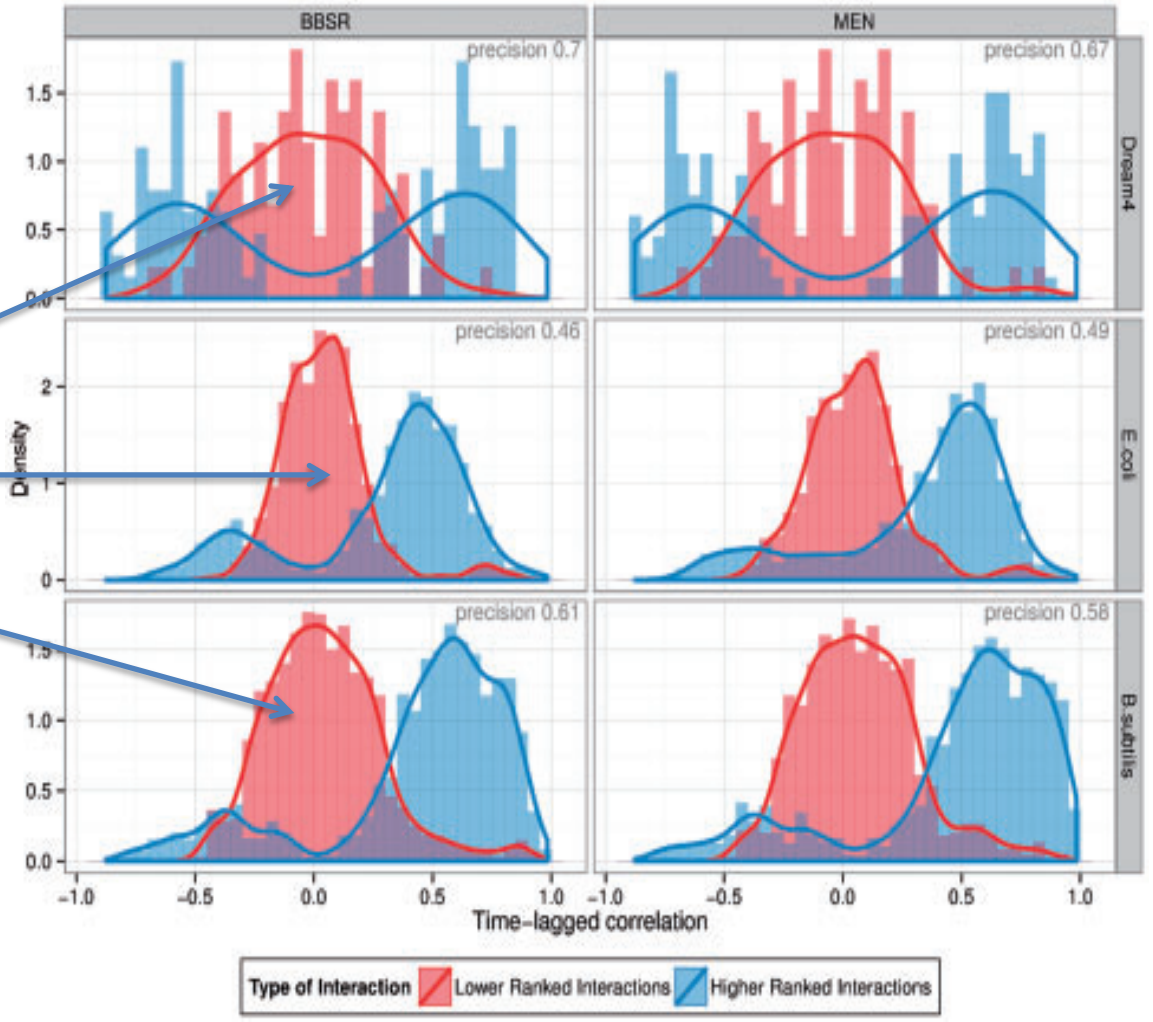
How does the prior parameter affect the performance?



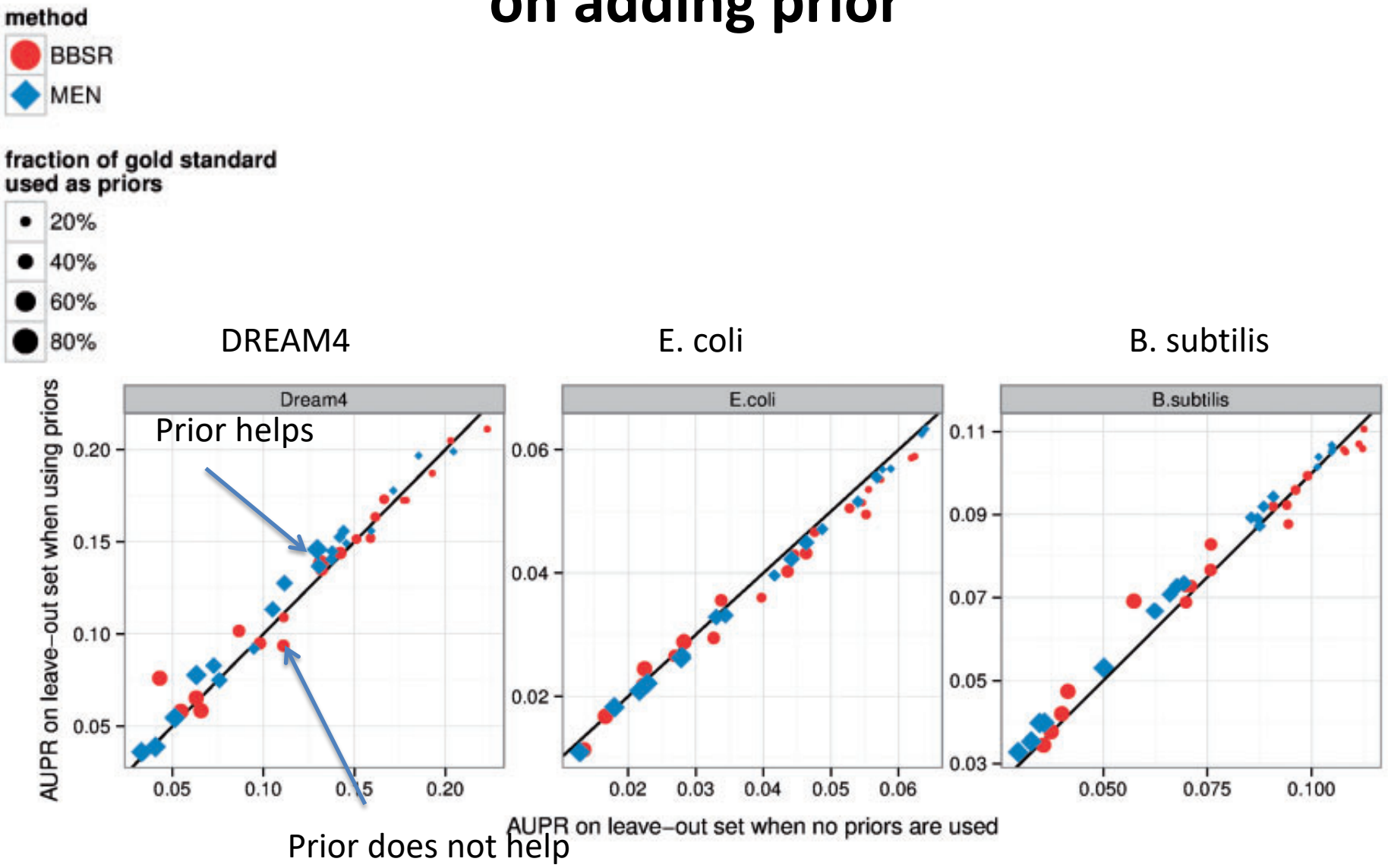
Can the data discriminate between different types of prior edges?

In other words, is the incorporation of prior data-driven?

Low ranked interactions do not have a strong positive or negative correlation



Ability to recover new edges is not hampered on adding prior



What happens when one adds noisy priors?

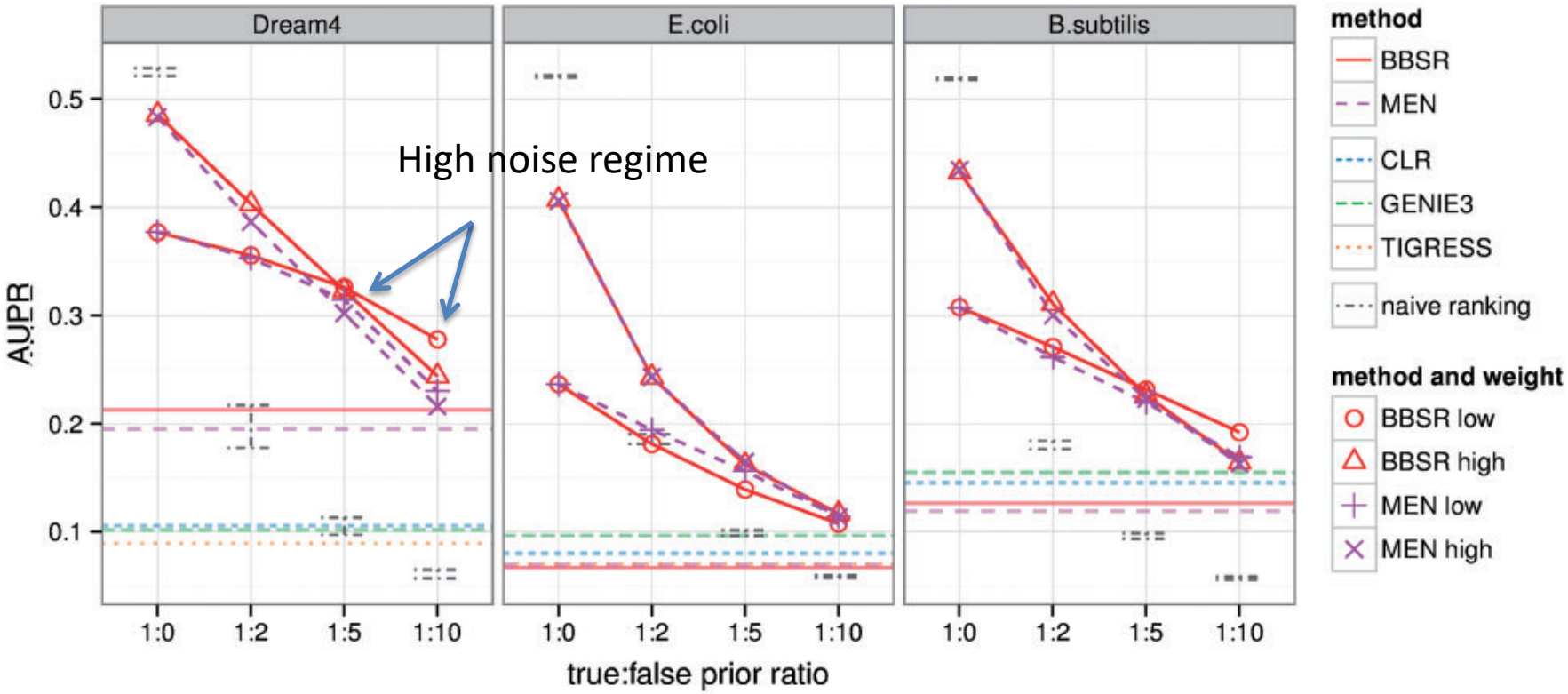


Fig. 5. Robustness to incorrect prior information. For each dataset, we considered half of the GSIs as TPIs, and added varying numbers of FPIs that were not GSIs. We show the AUPR of both methods for multiple choices of the respective weight parameters, as well as methods that do not use any PKIs (horizontal lines). Additionally, we show the performance of a naive interaction ranking method, which places all PKIs at the top of the list (gray bars)

Low and high in BBSR and MEN means less dense or more dense

Summary

- Extending the Inferelator linear regression model to incorporate priors
  - Regularized regression
  - Probabilistic priors on weights
- Experiments suggest
  - The prior incorporation is data-driven
  - Adding prior is beneficial even if when it is noisy

Goals for this lecture

- Incorporating priors in Dependency networks using linear regression
- Incorporating priors in Dependency networks using tree models

iRafNet

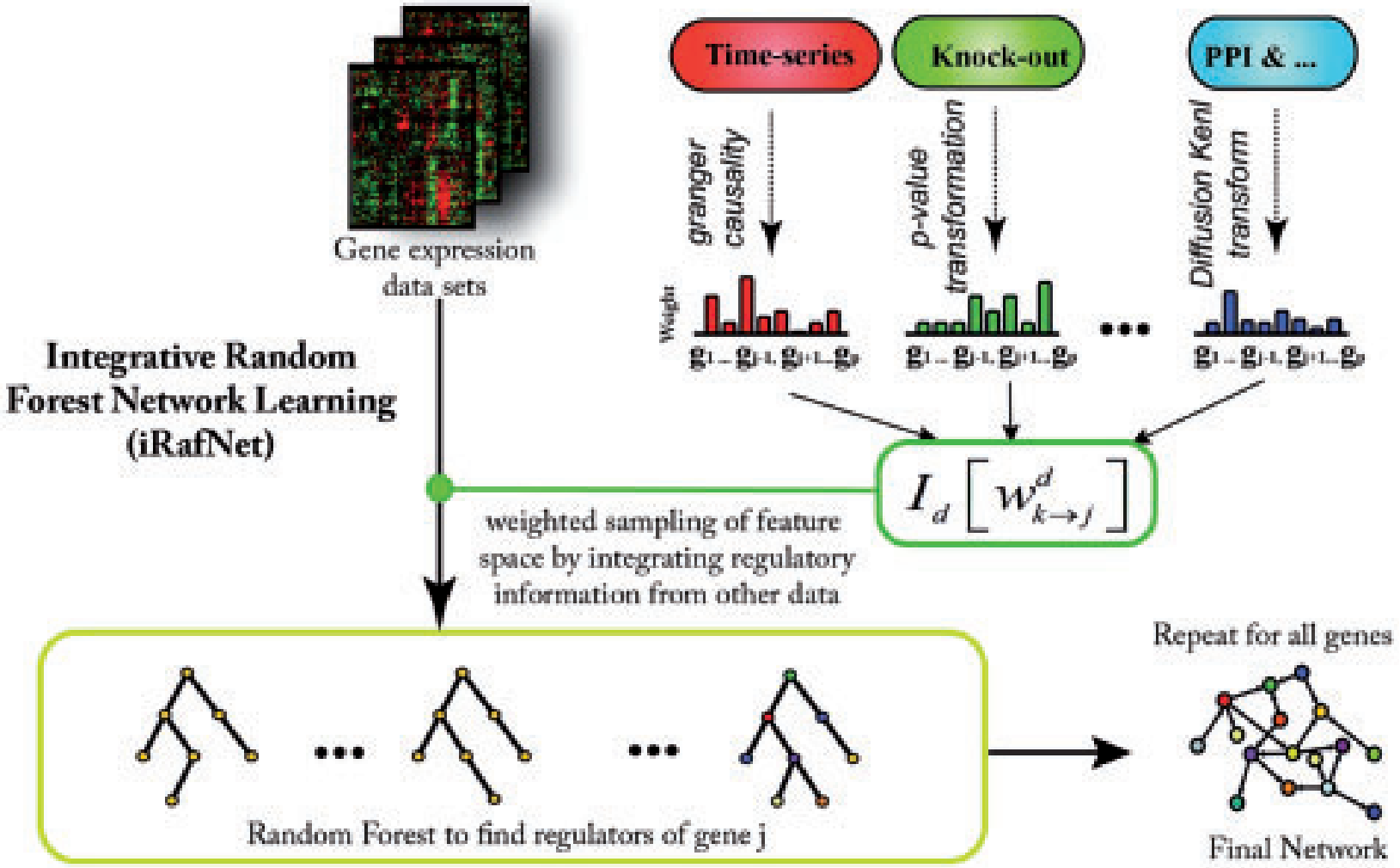
- GENIE3 was shown to be one of the best performing expression-based algorithms
- Can we extend the GENIE3 Random Forests based approach to incorporate priors?
- iRafNet uses a weighted sampling scheme to incorporate information from different sources of data

Petralia et al. 2015, Bioinformatics

Weighted sampling algorithm in iRafnet

- Each data source *d* provides a score for a regulator *k* and target *j*
- Convert these scores to sampling weights, $w_{k->j}$ in a data source and score-specific way
- For each node split, instead of sampling uniformly from *N* potential regulators, select a dataset *d* randomly and sample *N* regulators based on their weights in *d*

iRafNet overview



Petralia et al 2015, Bionformatics

Constructing sampling weights

- The prior knowledge is described as a set of weighted networks
- Weights for selecting a regulator is derived in a dataset specific manner
- Undirected protein-protein interactions:
  - Weights derived from a diffusion process over graphs (we will see this later lectures)
- Time-series expression data
  - Weight $w_{j->k}$ assess how predictive $\mathrm{g_j}$'s expression at time t is of $\mathrm{g_i}$'s expression at a future time point $t+1$
  - Derive a P-value to assess the strength of the regression weight
  - Convert P-value into a weight
- Knockout data
  - $w_{j->k}$ either are derived in multiple ways:
    - If g_k's expression changes significantly when g_k is knocked $w_{j->k}$ is derived from the P-value
    - Otherwise it is derived based on the overlap of g_j and g_k's knockout targets or knockout regulators

iRafNet application to real data

- Ground truth
  - Significant interactions identified from ChIP-chip experiments of yeast
- Expression dataset
  - This was a large study measuring gene expression in multiple yeast strains
- Prior datasets (included other expression datasets)
  - Expression time course during cell cycle
  - Expression data of genetic knockouts of TFs
  - Protein-protein interactions from public databases (BioGRID, MINT, DIP)

Does adding prior help for iRafNet?

Method	Data	AUC	AUPR
GENIE3	Expression	0.547 (0.537,0.566)	0.542 (0.537,0.548)
iRafNet	Multiple weights	0.624 (0.613,0.636)	0.565 (0.561,0.569)
	Expression and KO	0.657 (0.645,0.673)	0.567 (0.562,0.574)
	Expression and TS	0.543 (0.528,0.557)	0.536 (0.530,0.541)
	Expression and PPI	0.574 (0.562,0.591)	0.557 (0.551,0.561)

- Evaluate on ChIP-chip network of yeast
- Expression dataset
  - This was a large study measuring gene expression
- Prior datasets (included other expression datasets)
  - Expression time course during cell cycle
  - Knockout data from Hu et al
  - Protein-protein interactions from public databases (BioGRID, MINT, DIP)

Concluding remarks

- We have seen different ways to incorporate other data types to improve the quality of the inferred network
- Bayesian networks with structure prior
  - Use an energy function to assess concordance
  - Sensitive to incorrect prior information
- Dependency networks with priors
  - Linear regression approach aims to reduce the penalty on inferred edges
  - Tree-based approach enables a “biased” selection of regulators