Machine Learning Network-Constrained Regression of Epigenetic Data

Sivo Vladimirov Daskalov

Corpus Christi College

28 June 2017

Outline

- Epigenetic background
- Project goals
- Penalized regression methods
- Composite voting regression
- Orchestrated hyperparameter tuning
- Model evaluation
- Regression method similarities
- Breast cancer dataset

Epigenetic background

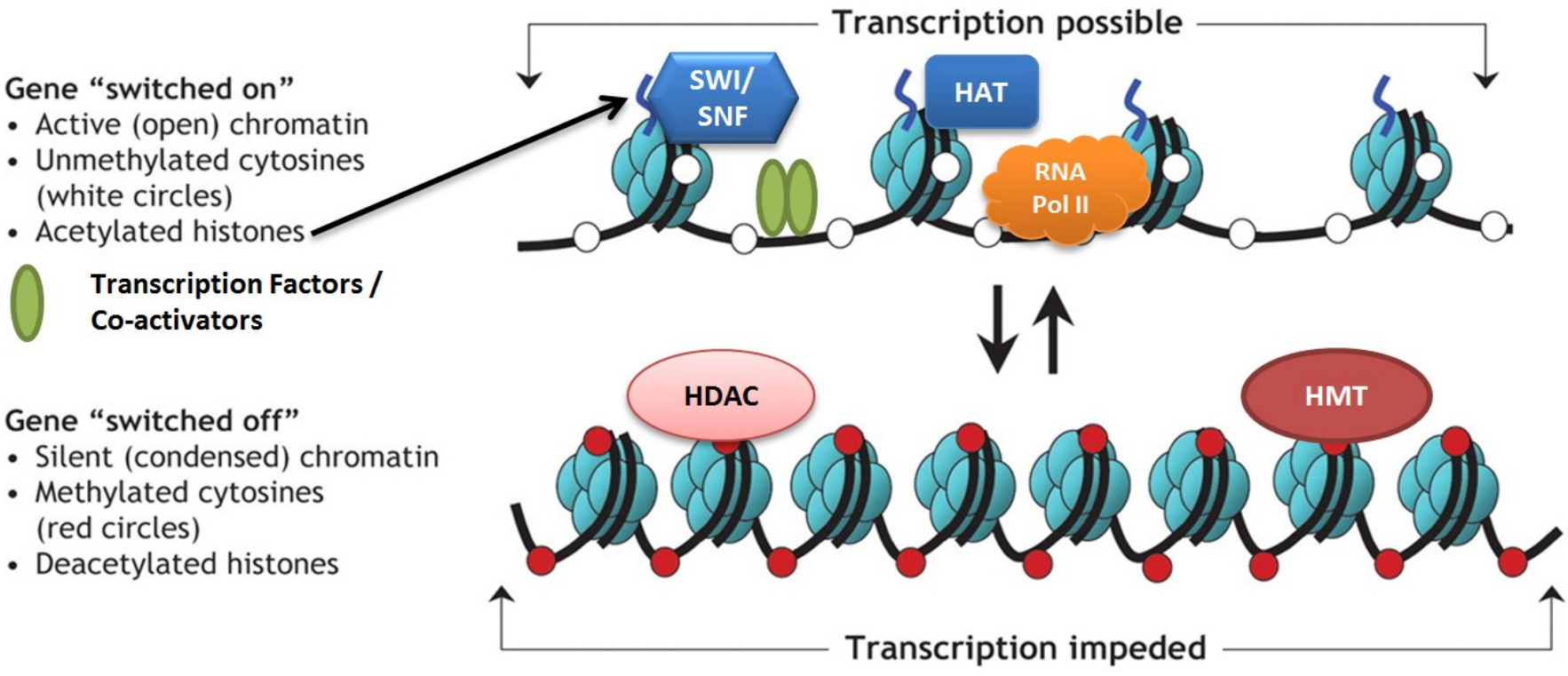


*Figure is adapted from Luong, P. Basic Principles of Genetics

Project goals

Question:
How is the expression of each gene affected by the methylation of related genes?

Approach:

Linear regression
- Predictors: methylation levels for all genes
- Target variable: expression level for gene of interest

Penalized regression methods

Lasso $\lambda \sum_{i=1}^{p} |\beta_i|$

Elastic Net $\lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sqrt{\sum_{i=1}^{p} \beta_i^2}$

Grace $\lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{u \sim v} \left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}} \right)^2 w(u,v)$

aGrace $\lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{u \sim v} \left(\frac{sign(\tilde{\beta}_u)\beta_u}{\sqrt{d_u}} - \frac{sign(\tilde{\beta}_v)\beta_v}{\sqrt{d_v}} \right)^2 w(u,v)$

GBLasso $\lambda \sum_{u \sim v} \left[\left(\frac{|\beta_u|}{\sqrt{d_u}} \right)^{\gamma} + \left(\frac{|\beta_v|}{\sqrt{d_v}} \right)^{\gamma} \right]^{1/\gamma}$

Linf $\lambda \sum_{u \sim v} \max \left(\frac{|\beta_u|}{\sqrt{d_u}}, \frac{|\beta_v|}{\sqrt{d_v}} \right)$

aLinf $\lambda \sum_{u \sim v} \left| \frac{sign(\tilde{\beta}_u)\beta_u}{\sqrt{d_u}} - \frac{sign(\tilde{\beta}_v)\beta_v}{\sqrt{d_v}} \right|$

TTLP $\lambda_1 \sum_{i=1}^{p} J_\tau |\beta_i| + \lambda_2 \sum_{u \sim v} \left| J_\tau \left(\frac{|\beta_u|}{w_u} \right) - J_\tau \left(\frac{|\beta_v|}{w_v} \right) \right|$

LTLP $\lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{u \sim v} \left| J_\tau \left(\frac{|\beta_u|}{w_u} \right) - J_\tau \left(\frac{|\beta_v|}{w_v} \right) \right|$

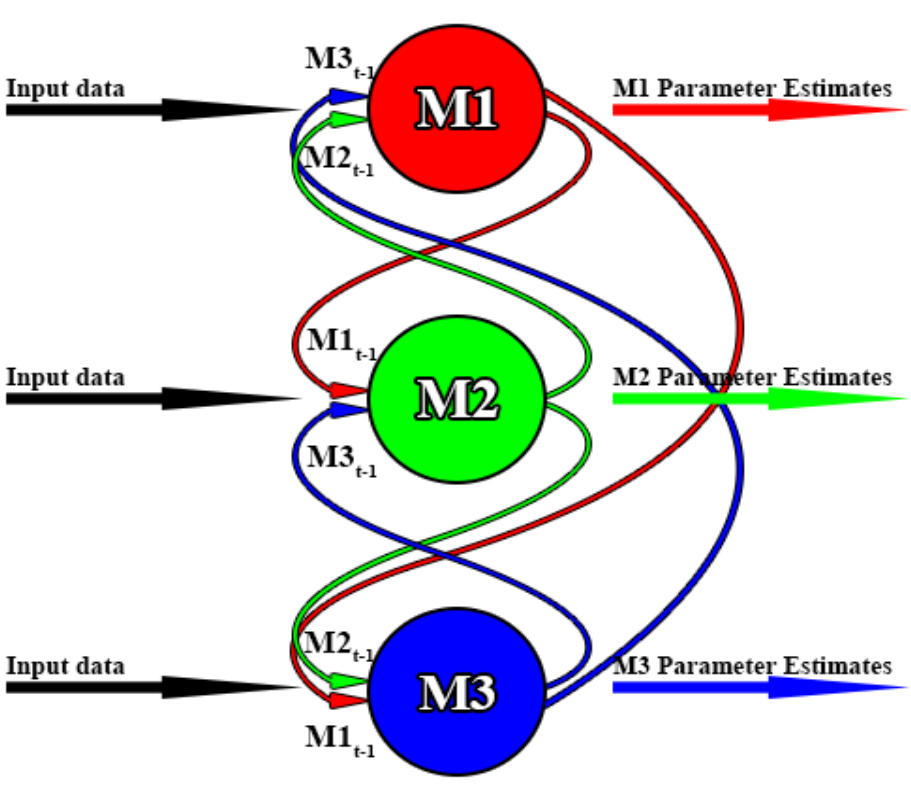
Composite voting regression

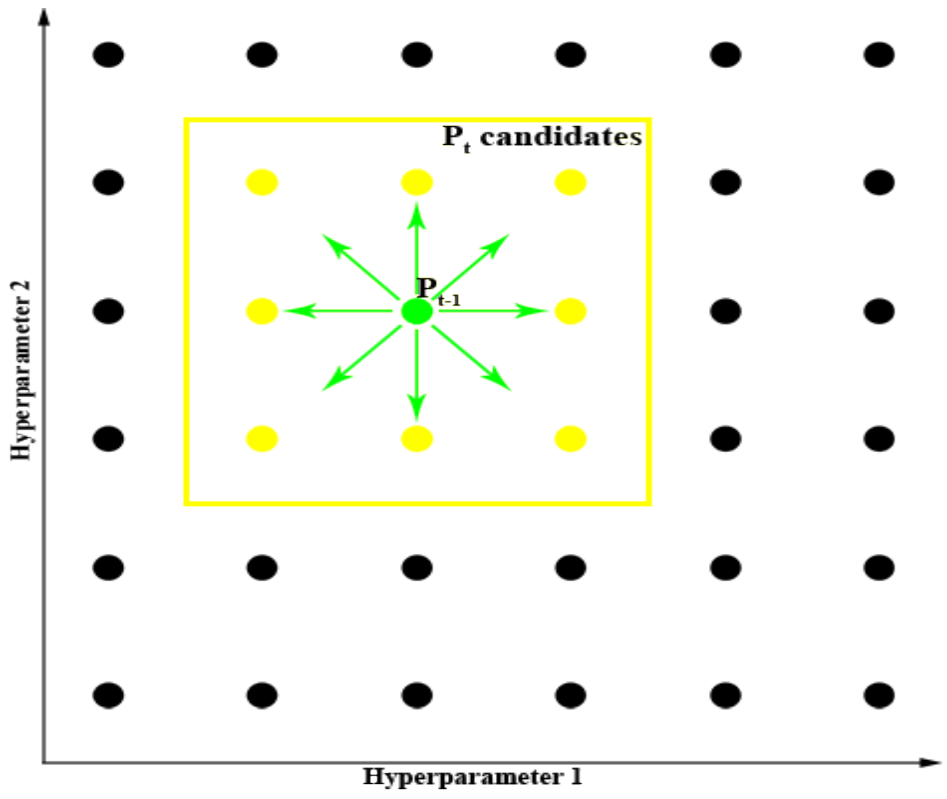
	X_1	X_2	...	X_p
Method 1	$M_1(\beta_1)$	$M_1(\beta_2)$	...	$M_1(\beta_p)$
Method 2	$M_2(\beta_1)$	$M_2(\beta_2)$	...	$M_2(\beta_p)$
...	...	...	...	...
Method k	$M_k(\beta_1)$	$M_k(\beta_2)$	...	$M_k(\beta_p)$

$$X_j = \begin{cases} \textit{important}, & \text{if } \dfrac{\sum_{i=1}^{k}[M_i(\beta_j) \neq 0]}{k} \geq \text{fraction of votes threshold} \\ \textit{unrelated}, & \text{otherwise} \end{cases}$$

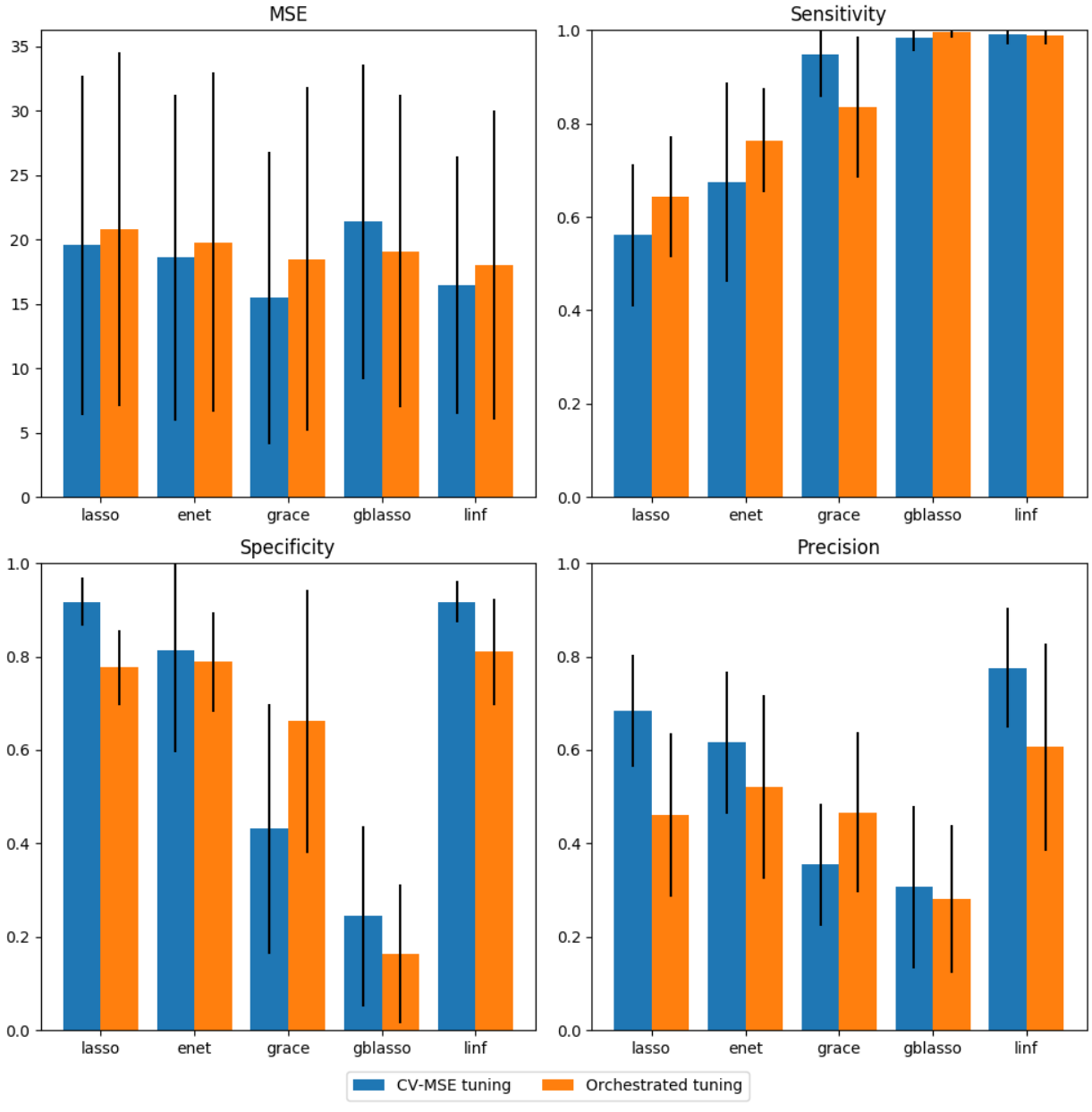
Final model obtained from OLSE on the set of important predictors

Orchestrated hyperparameter tuning

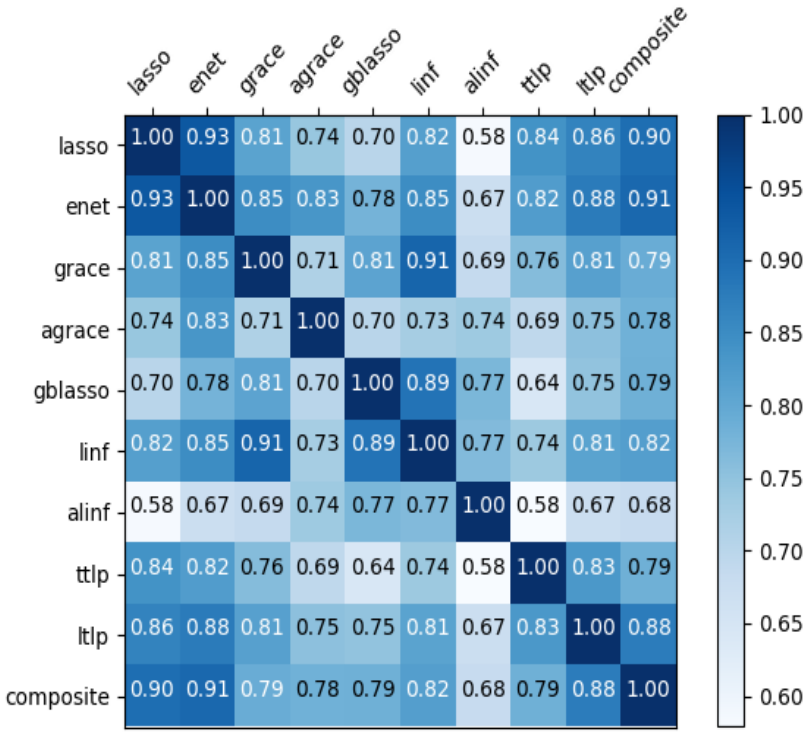




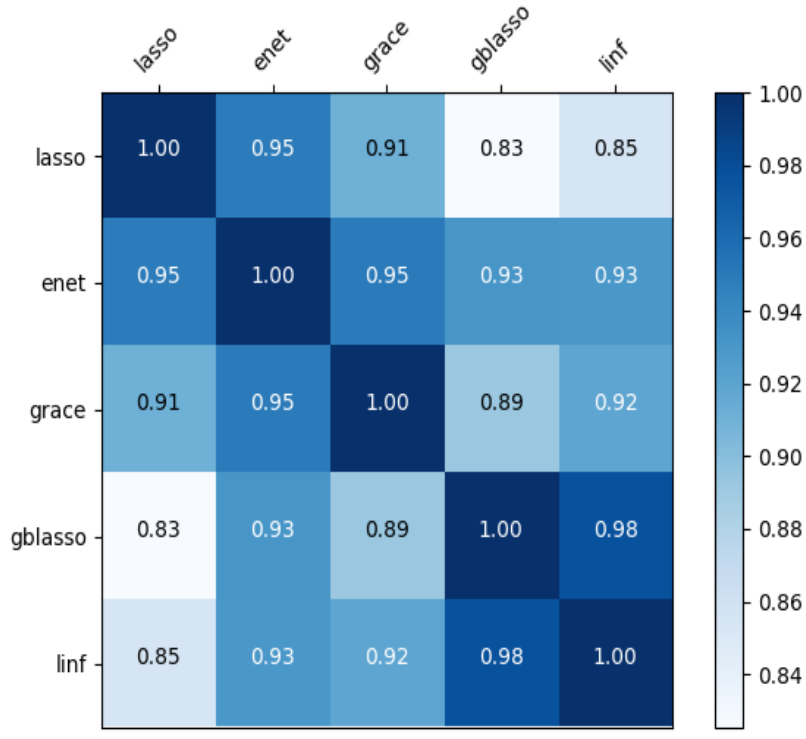
Model evaluation



Regression method similarities



CV-MSE tuning



Orchestrated tuning

Breast cancer dataset