Hub Gene Selection Methods for the Reconstruction of Transcription Networks

José Miguel Hernández-Lobato[1] and Tjeerd. M. H. Dijkstra[2]

(1) Computer Science Department, Universidad Autónoma de Madrid, Spain
(2) Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands

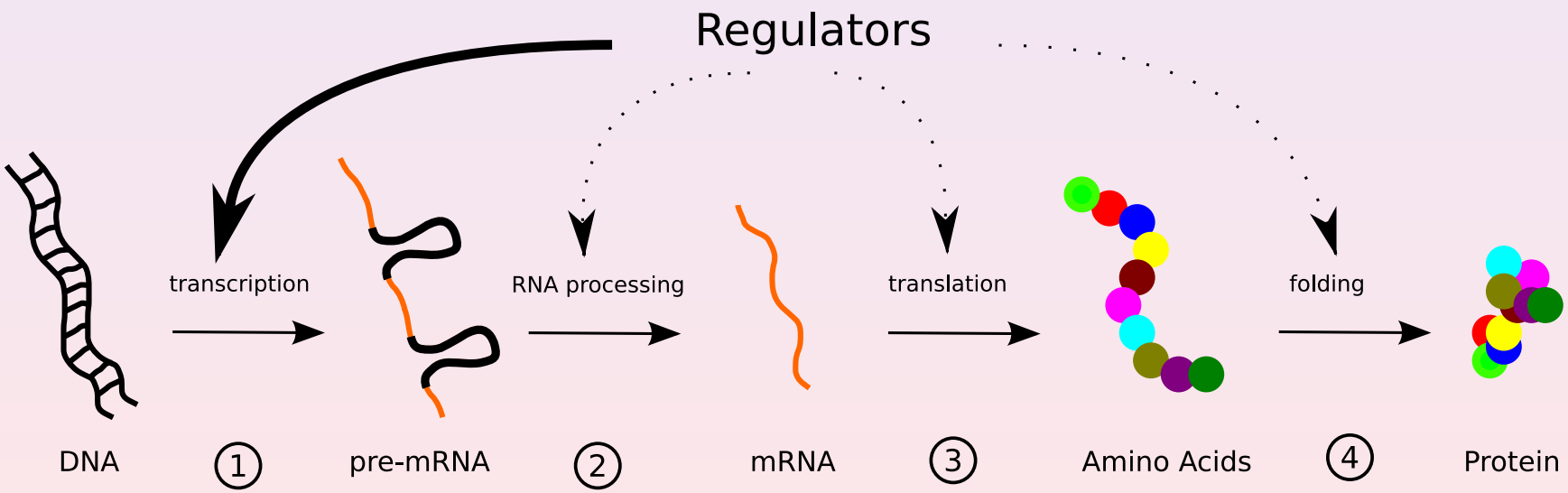
September 21, 2010

Outline

Outline

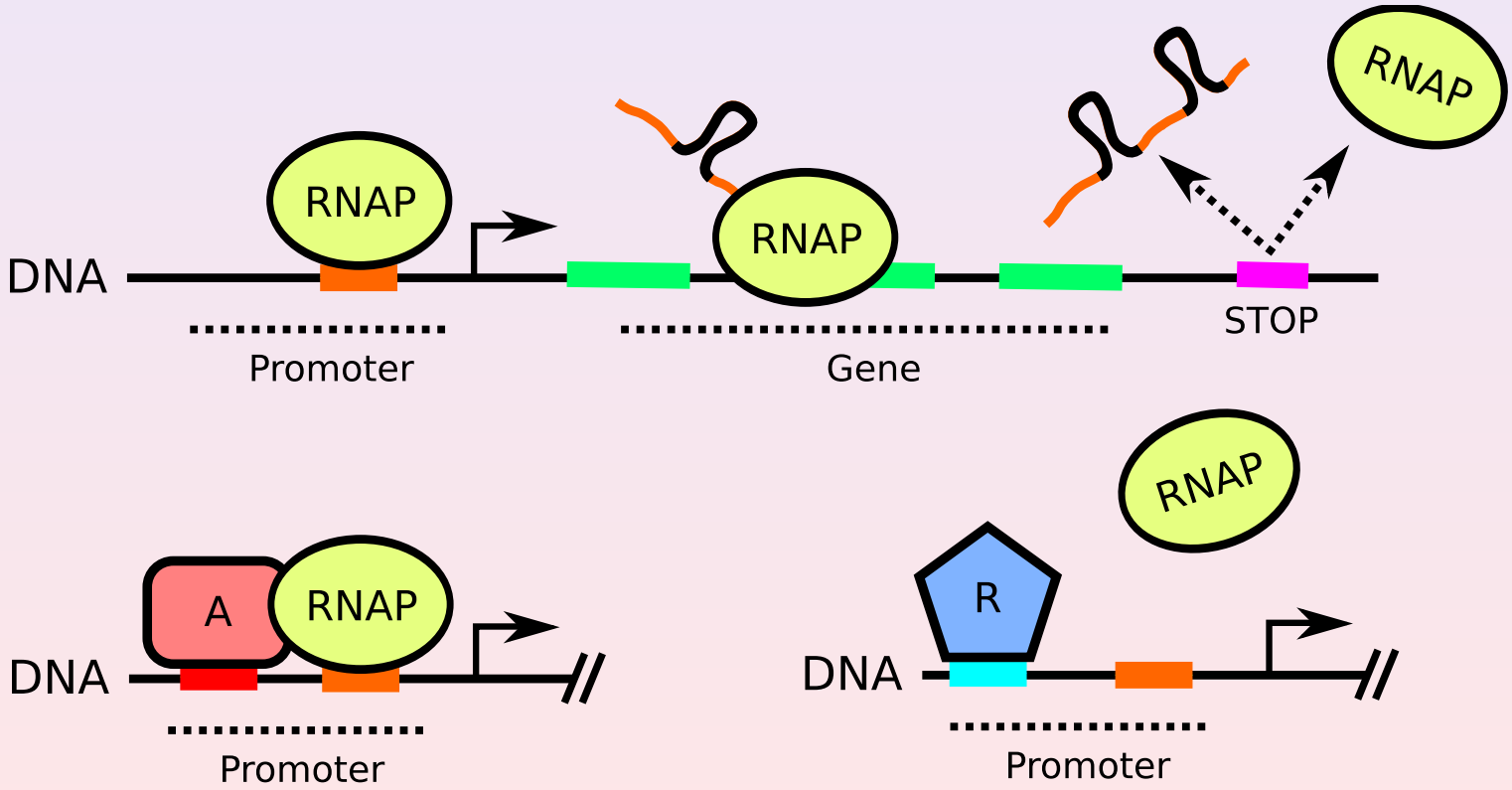
Gene Expression Process

Cells are composed of thousands of proteins produced at different rates in different situations. Regulatory molecules control the final concentration of each protein.



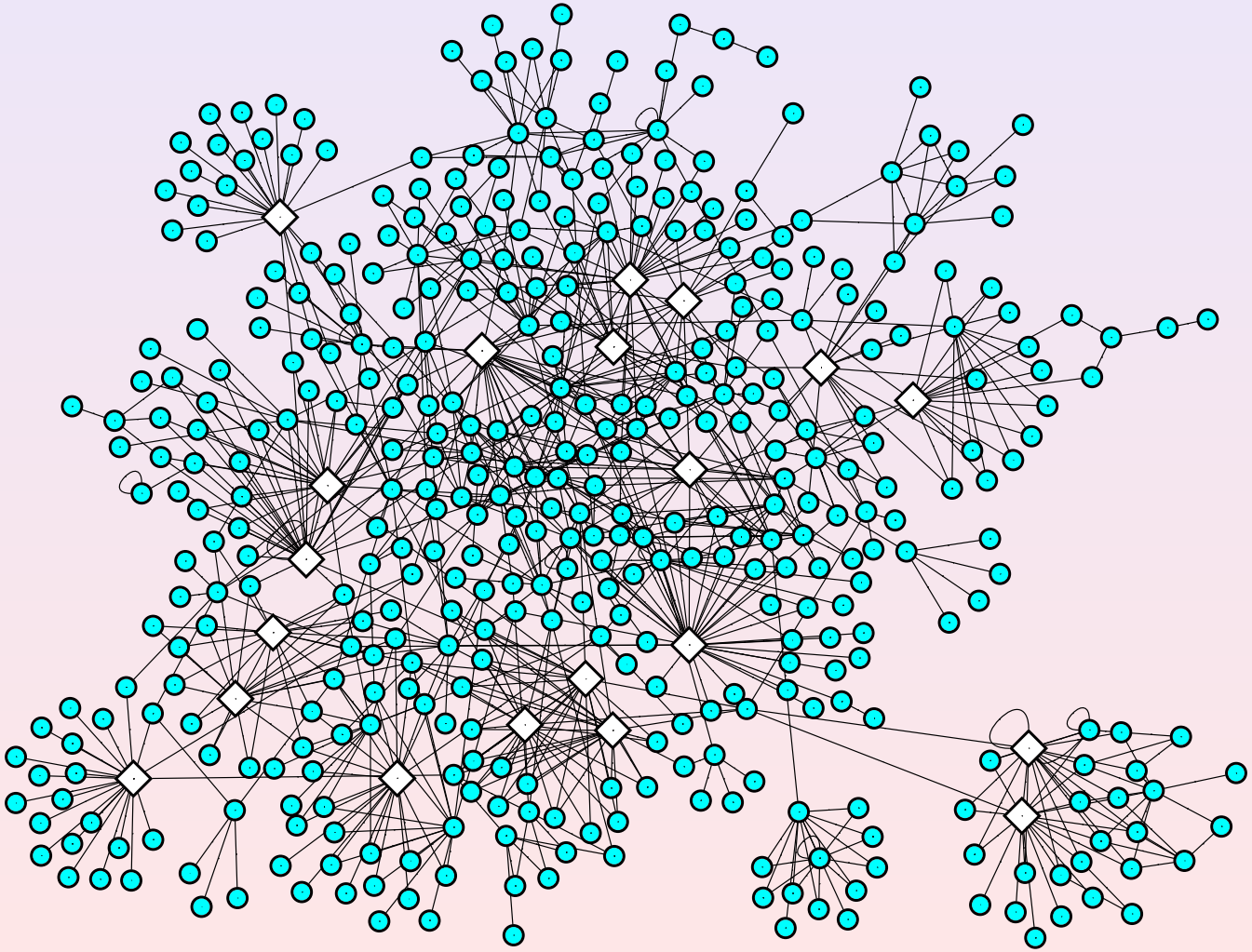
Transcription Regulation

Transcription is typically controlled by transcription factors which can be activators or repressors.

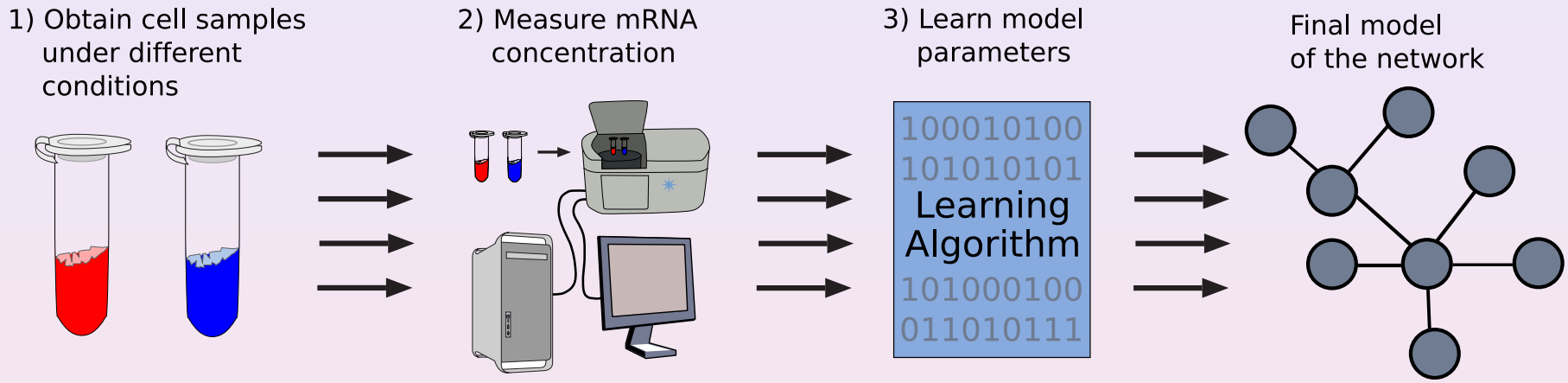


Transcription Control Networks...

...have hubs or highly connected genes. These are key regulators.



Reconstruction of Transcription Control Networks.



Main Contributions:

- Identifying beforehand those genes that are more likely to be hubs can improve the results of existing network reconstruction methods.
- We propose different methods for identifying hub genes.

Outline

A Linear Model for Transcription Control Networks

Under certain assumptions, the log-concentration of mRNA for a gene $[X_i]$ can be represented as a linear function of the log-concentration of mRNA for its regulators $[X_j]$ using a linear model:

$$\log[X_i] = \sum_{j \neq i} b_j \log[X_j] + \text{constant}\,.$$

where the b_j are regression coefficients. No autoregulation!

Extension to Account for all the Genes in a Cell

$$\mathbf{X} \approx \mathbf{B}\mathbf{X} + \sigma\mathbf{E}\,,$$

where $\mathbf{X}$ are mRNA log-concentration values and $\mathbf{E}$ is Gaussian noise. The diagonal of $\mathbf{B}$ is zero. Hub genes correspond to the columns of $\mathbf{B}$ that have more non-zero elements.

Basic Approach and Proposed Methods

Basic Approach

Identify the rows of $\mathbf{X}$ that are more relevant for solving the multiple linear regression problem $\mathbf{X} \approx \mathbf{B}\mathbf{X} + \sigma\mathbf{E}$, or the columns of $\mathbf{B}$ thar are more likely to have many non-zero elements. Each of these rows or columns represents a different hub gene.

Proposed Methods

We use standard feature selection and sparsity enforcing techniques:

- Automatic Relevance Determination (ARD).
- Group Lasso (GL).
- Maximum-relevance minimum-redundancy (MRMR).

Automatic Relevance Determination (ARD)

Let $\mathbf{b}_i^j$ be the element in the j-th column and in the i-th row of $\mathbf{B}$, then the prior for $\mathbf{B}$ is

$$\mathcal{P}(\mathbf{B}|\mathbf{a}) = \prod_i \delta(b_i^i) \prod_{j \neq i} \mathcal{N}(b_i^j, 0, a_j^{-1}), \tag{1}$$

where $\mathbf{a} = (a_1, \ldots, a_d)$ is a vector of inverse variances, one component for each column of $\mathbf{B}$ and δ is a point mass at zero. ARD maximizes the evidence of the resulting Bayesian model with respect to $\mathbf{a}$. During this process some of the a_i become infinite and the resulting solution for $\mathbf{B}$ is column-wise sparse. The optimization is performed in a greedy manner to select the k most important columns.

Group Lasso (GL)

A column-wise sparse estimate of $\mathbf{B}$ is obtained as the minimizer of

$$\| \mathbf{X} - \mathbf{B}\mathbf{X} \|_F^2 \quad \text{subject to} \quad \sum_{i=1}^{d} \| \mathbf{b}_i \|_2 \leq M \quad \text{and} \quad \text{diag}(\mathbf{B}) = \mathbf{0}\,, \tag{2}$$

where $\| \cdot \|_F$ and $\| \cdot \|_2$ stand for the Frobenius and ℓ_2 norms, respectively, $\mathbf{b}_i$ is the i-th column of matrix $\mathbf{B}$ and M is a positive regularization parameter. The ℓ_2 norm forces the solution to be column-wise sparse. M is fixed so that only k columns of the resulting estimate are different from zero.

Maximum Relevance Minimum Redundancy (MRMR)

Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$ be the i-th row of $\mathbf{X}$ and let $I(\mathbf{x}_i, \mathbf{x}_j)$ denote the empirical mutual information for rows i and j. Now, suppose that S is a set that contains those genes already selected by the method. The next gene to be selected is a gene j that is not yet in S and that maximizes

$$\frac{1}{d-1} \sum_{i \neq j} I(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{|S|} \sum_{i \in S} I(\mathbf{x}_i, \mathbf{x}_j) \,. \tag{3}$$

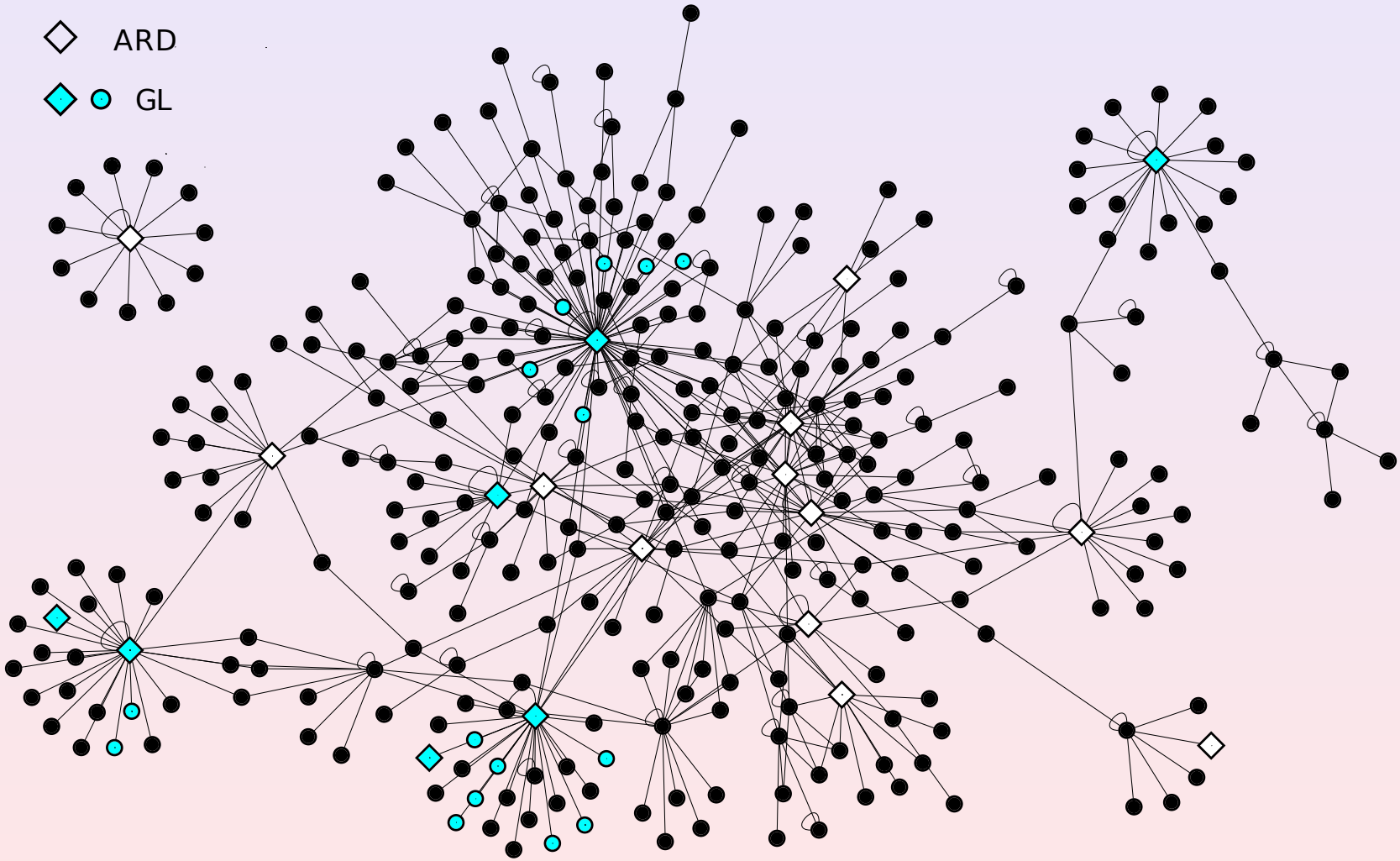
where $I(\mathbf{x}_i, \mathbf{x}_j) = -0.5 \log(1 - \hat{\rho}_{ij}^2)$, and $\hat{\rho}_{ij}$ is the empirical correlation of the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. The process is repeated until k genes are selected.

Evaluation of the Hub Gene Selection Methods

- SynTREN is used to generate synthetic gene expression data from three networks with 423, 690 and 1330 genes, respectively. 50 gene expression matrices with 250 measurements are generated from each network.
- The different methods are configured (fixing k) to select a 5% of genes.
- The methods are compared with a random approach that selects a 5% of genes randomly.
- Performance of each method is evaluated by the average connectivity of the selected genes.

Network	Genes	Random	GL	MRMR	ARD
Small E. coli	423	2.54±0.77	7.23±0.40	13.31±0.77	14.01±0.71
Yeast	690	3.14±0.91	9.46±0.51	13.91±1.27	16.51±0.74
Large E. coli	1330	4.22±1.94	8.56±0.72	14.70±2.48	23.48±1.62

Genes Selected on the Smallest Network by ARD and GL



Outline

Description of ARACNE

- ARACNE is a state-of-the-art method for TCN reconstruction.
- ARACNE follows the following steps:
  1. The empirical mutual information for any two genes i and j is computed, that is, $I(\mathbf{x}_i, \mathbf{x}_j)$.
  2. The connection weight w_{ij} for any two genes i and j is initialized as $w_{ij} = I(\mathbf{x}_i, \mathbf{x}_j)$.
  3. The data processing inequality is applied

$$I(\mathbf{x}_i, \mathbf{x}_j) \leq \min\{I(\mathbf{x}_i, \mathbf{x}_k), I(\mathbf{x}_j, \mathbf{x}_k)\} \tag{4}$$

     for any genes i, j and k and w_{ij} is set to zero to eliminate indirect interactions when the inequality holds for some gene k.
- Finally ARACNE links any two genes i and j when w_{ij} is higher than a threshold $\theta > 0$.

ARACNE and Noisy Data

- Noise in the mutual information estimates can significantly affect the performance of ARACNE.
- ARACNE may fail to remove an indirect interaction between genes i and j when fluctuations in the measurement process make $I(\mathbf{x}_i, \mathbf{x}_j)$ larger than $I(\mathbf{x}_i, \mathbf{x}_k)$ or $I(\mathbf{x}_j, \mathbf{x}_k)$ for those genes k that actually connect i and j.
- Similarly, ARACNE may mistakenly remove a direct interaction between genes i and j when these fluctuations make $I(\mathbf{x}_i, \mathbf{x}_j)$ smaller than $I(\mathbf{x}_i, \mathbf{x}_k)$ and $I(\mathbf{x}_j, \mathbf{x}_k)$ for some gene k.

Improving ARACNE using a Hub Gene Selection Method

After ARACNE has computed the empirical mutual information estimates, these are updated as

$$I_{\text{new}}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} I_{\text{old}}(\mathbf{x}_i, \mathbf{x}_j) + \Delta_{ij} & \text{if } i \in S_H \text{ or } j \in S_H \\ I_{\text{old}}(\mathbf{x}_i, \mathbf{x}_j) & \text{otherwise} \end{cases} \tag{5}$$

where S_H is a set of hub genes and Δ_{ij} is a positive number that scales with the level of noise in the estimation of $I(\mathbf{x}_i, \mathbf{x}_j)$.

When $I(\mathbf{x}_i, \mathbf{x}_j) = -0.5 \log(1 - \hat{\rho}_{ij}^2)$, we can fix a probabilistic upper bound on the noise in $I(\mathbf{x}_i, \mathbf{x}_j)$ using the delta method, namely

$$\Delta_{ij} = n^{-1/2} |\hat{\rho}_{ij}| \Phi^{-1}(1 - \gamma) , \tag{6}$$

where γ is a small positive number.

Evaluation of the New ARACNE Method I

- The standard ARACNE algorithm is compared with four versions of the modified method which compute S_H using ARD, GL, MRMR and the random approach
- The data used are the expression matrices generated by SynTREN to validate the hub gene selection methods.
- Performance is measured using the area under the precision-recall curve generated by altering threshold θ.
- γ is fixed to be the inverse of the number of genes in the network.

Network	Genes	Standard ARACNE	Random	New ARACNE with GL	MRMR	ARD
Small E. coli	423	0.41±0.03	0.28±0.05	0.41±0.04	0.57±0.04	0.56±0.03
Yeast	690	0.30±0.02	0.16±0.03	0.30±0.02	0.35±0.03	0.39±0.02
Large E. coli	1330	0.16±0.01	0.06±0.01	0.15±0.01	0.18±0.03	0.26±0.02

Outline

The ARD Method is Validated on a Yeast Dataset...

...which includes 247 expression measurements for a total of 5520 genes and it is publicly available at the Many Microbe Microarrays Database.

Table: Top ten genes selected by ARD on the yeast dataset.

Rank	Gene	Description
1	YOR224C	RNA polymerase subunit.
2	YPL013C	Mitochondrial ribosomal protein.
3	YGL245W	Glutamyl-tRNA synthetase.
4	YPL012W	Ribosomal RNA processing.
5	YER125W	Ubiquitin-protein ligase.
6	YER092W	Associates with the INO80 chromatin remodeling complex
7	YBR289W	Subunit of the SWI/SNF chromatin remodeling complex.
8	YBR272C	Involved in DNA mismatch repair.
9	YBR160W	Catalytic subunit of the main cell cycle kinase.
10	YOR215C	Unknown function.

Outline

Conclusions

- Network reconstruction can be improved by predicting beforehand which genes are hubs.
- The linear model identifies hubs with non-sparse columns of a regression matrix **B**.
- We have proposed three hub gene selection methods: ARD, GL, MRMR. The best one is ARD.
- The performance of ARACNE can be improved by using some of these methods.
- The best method, ARD, was validated on a yest expression dataset.

Bibliography

- Gardner, T.S., Faith, J.J.: Reverse-engineering transcription control networks. Physics of Life Reviews 2(1), 65-88 (2005)
- den Bulcke, T.V., Leemput, K.V., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., Moor, B.D., Marchal, K.: Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics 7(1), 43 (2006)
- Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A.: Reverse engineering of regulatory networks in human B cells. Nature Genetics 37, 382-390 (2005)
- Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 4967 (2006)
- Tipping, M.E.: Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 1, 211-244 (2001)
- Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226-1238 (2005)
- Faith, J.J., Driscoll, M.E., Fusaro, V.A., Cosgrove, E.J., Hayete, B., Juhn, F.S., Schneider, S.J., Gardner, T.S.: Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. Nucleic Acids Research 36, D866-D870 (2008)
- D.P. Wipf and B.D. Rao, An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem, IEEE Transactions on Signal Processing, vol. 55, no. 7, July 2007

Thank you for your attention!