Maximum Likelihood Quantization and Logical Networks for Modeling Biological Interactions

Mingzhou (Joe) Song[1], Eric R. Lance[2], Chris K. Lewis[3], Elissa J. Chesler[4], Roumyana Kirova[5], Susan E. Bergeson[6]

Keywords: Quantization, logical networks, gene regulatory networks, response to alcohol

1   Problem statement.

We address the computational problems of discretizing continuous random variables and identifying a logical network model to account for temporal dependencies among genes and environmental stimuli from high-throughput transcriptome data. These new algorithms were applied to real biological data from an analysis of the molecular response to alcohol. Our approach captures statistical interactions in the quantization step, thereby retaining a high level of information from the input data. This results in minimal loss of signal from data reduction for logical network construction, in contrast to conventional methods which discretizes along a single dimension–resulting in an information poor approach to input variable selection.

2   Results.

We proved a theorem that establishes a theoretical foundation for the design of quantization algorithms so that interactions among original variables are not lost after discretization. Under our theoretical argument, we designed an approximation algorithm Quantization-by-Deletion, with time complexity bounded between $\Omega(d^2 n^3 \log n)$ and $O(d^3 n^3 \log n)$ where d is dimension and n is sample size, to preserve statistical conditional independence among original continuous random variables. On the quantized data, we developed a logical network construction algorithm that uses the statistical significance as a criterion for network selection to avoid false interactions arising from pure chance. Using temporal gene expression data collected from the brains of alcohol treated mice, our algorithms identified several genes from a major neural pathway as components of the alcohol response. Three genes have specific associations with alcohol response as reported in the literature. We also highlighted several other potential genes, agreeing with independent literature mining results [1, 3, 6], that could play a role in response to alcohol. In addition, unknown interactions discovered in the logical network, subject to biological verification, will offer new clues for biologists to study the elusive molecular mechanisms of alcoholism.

[1]Department of Computer Science, New Mexico State University, NM, USA. E-mail: `joemsong@cs.nmsu.edu`

[2]Klipsch School of Electrical and Computer Engineering, New Mexico State University, NM, USA. E-mail: `elance@nmsu.edu`

[3]Department of Computer Science, New Mexico State University, NM, USA. E-mail: `chrislew@nmsu.edu`

[4]Biosciences Division, Oak Ridge National Lab, TN, USA. E-mail: `cheslerej@ornl.gov`

[5]Biosciences Division, Oak Ridge National Lab, TN, USA. E-mail: `yordanovark@ornl.gov`

[6]School of Biological Sciences, University of Texas at Austin, TX, USA. E-mail: `bergeson@mail.utexas.edu`

3   Significance.

Our joint quantization algorithm converts expressions of hundreds of genes into discrete values while retaining information about higher order interactions influencing gene expression regulation. In this process, only genes that are strongly interacting with others receive more than one quantization level. A gene exhibiting marginal variation but no interaction with other genes is unlikely to be discretized into more than one quantization level. The quantization algorithm achieves this desirable effect by maximally preserving statistical conditional independence among original continuous random variables. This approach specifically enriches model input to detect gene-gene interactions. The concept of conditional independence is incorporated directly into the quantization strategy to allow discretization of random variables without destroying the interaction structure in the original data. The multinomial test based logical network construction algorithm systematically assesses the statistical significance of a logical network to avoid obtaining models out of pure chance and constitutes a new approach to building logical networks. This biological significance is reflected in our analysis of alcohol effects on brain gene transcription, for which we will conduct biological validation experiments.

4   Comparison to other approaches.

Joint quantization is still not the mainstream due to its often high computational cost, but it is being increasingly adopted, as the computationally less intensive single dimensional quantization is not designed to capture interactions among random variables. Our quantization approach, a development from [5], maximizes probability density based likelihood, different from all alternative approaches such as [2], is a major advantage in that it preserves interactions among the original continuous random variables. The statistical significance assessment of a logical network is another important advancement from previous approaches such as [4] which often do not evaluate the chance of a model arising randomly from noninteracting random variables. The data we have chosen for this study come from one of the first to explore temporal patterns of gene expression following alcohol administration.

References

[1] Bachtell, R. K., Wang, Y. M., Freeman, P., Risinger, F. O. and Ryabinin, A. E. 1999. Alcohol drinking produces brain region-selective changes in expression of inducible transcription factors. *Brain Res.*, 847(2): 157–165.

[2] Friedman, N. and Goldszmidt, M. 1996. Discretizing continuous attributes while learning Bayesian networks. In *Proc. of Int'l Conference on Machine Learning*, pages 157–165.

[3] Ozimek, P., Lahtchev, K., Kiel, J. A., Veenhuis, M. and van der Klei, I. J. 2004. Hansenula polymorpha Swi1p and Snf2p are essential for methanol utilisation. *FEMS Yeast Res.*, 4 (7): 673–682.

[4] Shmulevich, I., Dougherty, E. R., Kim, S. and Zhang, W. 2002. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18: 261–274.

[5] Song, M. and Haralick R. M. 2002. Optimal Multidimensional Quantization for Pattern Recognition. In *Proc. of SPIE Vol. 4875: 2nd Int'l Conf. on Image and Graphics*, volume Part 1, pages 15–30.

[6] Winston, N. J. and Maro, B. 1995. Calmodulin-dependent protein kinase II is activated transiently in ethanol stimulated mouse oocytes. *Dev Biol.*, 170(2):350–352.