The Dynamic and Modular Transcriptome of an Extremophilic Archaean

Peter Tonner* peter.tonner@duke.edu
Barbara Engelhardt†
Amy Schmid‡

ABSTRACT

Hypersaline adapted Archaea, due to their habitation of harsh desert salt lakes, can withstand stress at extreme levels. Their survival in extreme stress relies on genome-wide shifts in transcription, but it is unclear how their transcriptome is organized locally and globally to ensure survival. We leverage over a thousand microarrays for the model extremophile *Halobacterium salinarum* to understand the dynamic states of its transcriptome. Two machine learning models allow us to decompose *H. salinarum* gene expression into sets of co-regulated modules. We investigate the genomic structures of these modules and use them to predict other phenotypes.

Keywords

Arhcaea, Transcriptional Regulation, Gaussian Graphical Models, Stochastic Block Models, Extreme stress response

H. salinarum resists numerous stress conditions at levels higher than other organisms, for example tolerating levels of reactive oxygen species 25 times higher than *E. coli* [1]. *H. salinarum* and the halophilic Archaea are also known to be diverse in their metabolic function, being able to perform aerobic respiration, fermentation of arginine, and phototrophy through the light activated proton pump *bacteriorhodopsin*. Both stress response and metabolism are regulated heavily at the transcriptional level, making transcriptional regulation an important area of systems biology in Archaea. Transcription regulation in Archaea is a hybrid of Eukaryotes and Bacteria; the general transriptional machinery resembles those of Eukaryotes and the transcriptional regulators resemble those of Bacteria (on the amino acid sequence level).The organization of *H. salinarum*'s transcriptional network, which allows the organism to survive diverse sources of and extreme levels of stress, is still unknown.

H. salinarum represents the model for Archaeal transcriptional regulation as it was adopted early for full genome sequencing and subsequent gene expression microarray analysis [2]. This has lead to over a decade of gene expression research for this organism, representing one of the largest collections of gene expression for a single organism. The protocol is consistent between all experiments, using the same reference strain in all arrays, which reduces the statistical variability inherent in most gene expression studies across time and labs.

We apply two popular machine learning algorithms to the *H. salinarum* transcriptional data — Gaussian graphical models and stochastic block models. A Gaussian graphical model (GGM) is a graphical (network) representation of a multivariate normal distribution, where the graph G is the inverse of the covariance matrix. An edge in G ($G_{i,j} \neq 0$) means the two nodes i and j are correlated when conditioned on all other variables.

Stochastic block models (SBMs) describe a network as a collection of k clusters, with the probability of an edge between two nodes described by their cluster memberships. The hidden cluster memberships of each node ($\pi_i \in \mathbb{R}^k$) and inferred block structure $B \in \mathbb{R}^{k \times k}$ describing the probability of edges between two clusters can fully factorize the network. The probability of each edge can then be represented as $p(G_{i,j}) = \pi_i^T B \pi_j$.

We use a sparse GGM [3] and a mixed membership SBM [4] to infer the complex transcriptome of *H. salinarum*. Module structures found from these models are then analyzed for their specificity of condition activity, overrepresentation of function, and conserved promoter motifs.

References

[1] Kriti Sharma et al. "The RosR transcription factor is required for gene expression dynamics in response to extreme oxidative stress in a hypersaline-adapted archaeon". en. In: *BMC Genomics* 13.1 (July 2012), p. 351.

[2] W. V. Ng et al. "Genome sequence of Halobacterium species NRC-1". en. In: *Proceedings of the National Academy of Sciences* 97.22 (Oct. 2000), pp. 12176–12181.

[3] Cho-Jui Hsieh et al. "Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation". In: *arXiv:1306.3212 [cs, stat]* (June 2013).

[4] Edoardo M. Airoldi et al. "Mixed Membership Stochastic Blockmodels". In: *J. Mach. Learn. Res.* 9 (June 2008).

*Computational Biology and Bioinformatics Program, Duke University
†Biostatistics, Duke University
‡Biology, Center for Systems Biology, Duke University