Practical Application of a Bayesian Network Approach to Poultry Epigenetics and Stress

Emiliano Ariel Videla Rodriguez[1] **Fábio Pértille**[2] **John B.O. Mitchell**[3] **Per Jensen**[2]
Carlos Guerrero-Bosagna[2] **V. Anne Smith**[1]

[1]School of Biology, University of St Andrews, St Andrews, Fife KY16 9TH, United Kingdom
[2]AVIAN Behavioural Genomics and Physiology Group, Department of Physics, Chemistry and Biology, Linköping University, 58183, Sweden
[3]School of Chemistry, University of St Andrews, St Andrews, Fife KY16 9ST, United Kingdom

1 EXTENDED ABSTRACT

Probabilistic networks can explore and describe the relationships among genetic or epigenetic features, even allowing the possibility to consider a particular condition of interest as part of the network (e.g., a stressful condition) [Agrahari et al., 2018, Jansen et al., 2003]. Bayesian networks represent a useful approach that has been applied to many biological systems in order to model the dependencies among a given set of features [Heckerman et al., 1995]. The aim of our study was to apply a Bayesian network structure learning approach in order to elucidate relationships and potential interactions among epigenetic features and a stressful condition driving epigenetic differences, in a poultry animal model, the chicken (*Gallus gallus*).

The epigenetic data presented challenges for learning the Bayesian network structure. Firstly, while the 46 animals in the study represent a relatively large cohort in epigenetics, it is a reasonably small quantity of data for Bayesian networks, resulting in a challenging search space. Secondly, the epigenetic data had an imbalance in discrete states, which can lead to possible artefacts in learning the structure [Milns et al., 2010, Mitchell et al., 2021]. In order to overcome these challenges, we applied an interdisciplinary approach by using advances first developed for Bayesian networks within an ecological context [Milns et al., 2010] to our poultry epigenetics dataset.

A Bayesian network structure was learned from epigenetic data collected from our experiment involving 46 male White Leghorn chickens (*Gallus gallus*). Twenty-two chickens were raised under control conditions, while the other twenty-four were exposed to a social isolation protocol for 21 consecutive days (stressful condition) [Pértille et al., 2020]. Thereafter, the DNA was extracted from red blood cells, a reduced representation of the methylome was sequenced, and bioinformatics pre-processing and analysis was performed. A set of 60 regions were selected, each one of them having differential methylation patterns of the nucleotide cytosine (henceforth referred to as differentially methylated regions or DMRs) between the experimental groups. Each individual had a particular value for each DMR, depending on the number of DNA segments containing methyl groups added to cytosine, ranging from 0 to 39 [Pértille et al., 2020]. In addition to the DMRs, the stressful condition was included in the dataset as a binary variable (control $= 0$; stress $= 1$).

Analysis of the DMR distributions showed that non-methylation (DMR $= 0$) was the most abundant state. Thus, the values of each DMR were also discretized into two possible categories, 0 for absence of methylation, and 1 for presence of methylation, the latter including all values > 0. Despite this discretisation, there was still an overabundance of samples without methylation, creating an imbalance of 0s and 1s. To avoid potential artefacts from this imbalance, we applied the methodology from Milns et al. [2010]. A chi-square test was applied for each pair of variables, filtering p-values equal or higher than 0.25. These pairs were included as prior information as arcs to be avoided in the structure, as the test showed no possible dependence between them.

A total of 100 Bayesian networks were learnt from an initial set of 100 random graph, using the R package `bnlearn` with a tabu search and the BDe score [Scutari, 2011]. To deal with the challenging search space, the phylogenetic model averaging approach from Milns et al. [2010] was used to select high probability arcs. This phylogenetic approach treats the presence of arcs in a network as features with which to build a phylogenetic tree, then performs a regression on network score controlled by the correlation patterns of the phylogenetic tree in order to identify an average probability of each arc being in a high-scoring network. These arc probabilities are clustered using a Gaussian mixed model to identify highly probable arcs [Milns et al., 2010]. Considering that this was still a heuristic random process, the final set of highly probable arcs identified was slightly different after different runs of the algorithm. Therefore, in order to build the consensus network, the arcs common to 50 searches were combined. The Markov Blanket property of the stressful condition was identified as the set of parents,

Accepted for 15th Bayesian Modeling Applications Workshop at UAI 2021 (BMAW2021).

children and spouse nodes that makes the stressful condition independent from the rest of the network.

The consensus network included 47 out of the 61 features, identifying a total of 43 arcs. OCLN—DMR7, CANX—TPST2, and FBN1—ENSGALG00000027231.4 had the highest values of probabilities (0.96, 0.86 and 0.83, respectively). The Markov Blanket of the stressful condition consisted of two DMRs, OCLN and ENSGALG00000051236.1. The arc between the treatment and OCLN had the highest average probability value (0.81).

Bayesian networks are a useful tool to identify and unravel hidden patterns within the data. For example, identification of methylation patterns of as-yet unnamed genes (ENSGALG00000051236.1 and ENSGALG00000027231.4) as being relevant in the network can spur biological investigation into the function of these genes. In particular for genetic and epigenetic networks, the focus can be put on the overall structure of the network and how the information flows through it [Li et al., 2010]. The Markov Blanket of the stressful condition could be used as epigenetic markers in close association with stress resilience, working towards the identification of biomarkers to be used for animal welfare improvements.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812777.

References

Rupesh Agrahari, Amir Foroushani, T. Roderick Docking, Linda Chang, Gerben Duns, Monika Hudoba, Aly Karsan, and Habil Zare. Applications of Bayesian network models in predicting types of hematological malignancies. *Scientific Reports*, 8:6951, 2018.

David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J. Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F. Greenblatt, and Mark Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302: 449–453, 2003.

H Li, G Wu, J Zhang, and N Yang. Identification of the heart-type fatty acid-binding protein as a major gene for chicken fatty acid metabolism by Bayesian network analysis. *Poultry Science*, 89:1825–1833, 2010.

Isobel Milns, Colin M. Beale, and V. Anne Smith. Revealing ecological networks using Bayesian network inference algorithms. *Ecology*, 91:1892–1899, 2010.

Emily G. Mitchell, Margaret I. Wallace, V. Anne Smith, Amanda A. Wiesenthal, and Andrew S. Brierley. Bayesian Network Analysis reveals resilience of the jellyfish *Aurelia aurita* to an Irish Sea regime shift. *Scientific Reports*, 11:3707, 2021.

Fábio Pértille, Adriana Mercia Guaratini Ibelli, Maj El Sharif, Mirele Daiana Poleti, Anna Sophie Fröhlich, Shiva Rezaei, Mônica Corrêa Ledur, Per Jensen, Carlos Guerrero-Bosagna, and Luiz Lehmann Coutinho. Putative epigenetic biomarkers of stress in red blood cells of chickens reared across different biomes. *Frontiers in Genetics*, 11:508809, 2020.

Marco Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35:1–22, 2011.