Extending Similarity Network-Based Classifiers to the Non-Coding Genome and Deep Learning

Jennifer Yu
University of Waterloo
Waterloo, Canada
jennifer.yu1@uwaterloo.ca

Duncan Forster
Donnelly Centre for Cellular
and Biomolecular Research
Toronto, Canada
duncan.forster@mail.utoronto.ca

Shraddha Pai
Ontario Institute for Cancer Research
Toronto, Canada
shraddha.pai@utoronto.ca

Summary

Similarity networks provide a useful framework for multi-modal data integration, suitable for applications such as gene function prediction and patient classification (Wang et al. [2014], Pai and Bader [2018]). We previously developed a supervised learning algorithm which converted heterogeneous patient data into the common space of patient similarity networks (PSN) and used these networks as input features (Pai et al. [2019]; netDx.org). In addition to excellent classification performance and handling missing data, netDx provides interpretability by allowing users to group genes into pathway-level features. However, the pathway-based grouping approach is of limited value for genomic data outside coding regions. Moreover, the current framework has limited scalability in the number of nodes and networks and does not take advantage of improved discriminability available in the deep learning framework. Here, we describe two recent areas of work addressing these limitations. In the first, we classify binary survival in PFA ependymomas using DNA methylomes organized using prior knowledge of brain tissue- and cell-specific expression, transcription factor binding sites and chromatin state. In the second, we extend a recently developed framework from Forster et al. [2021] for multiple network integration based on graph convolutional networks, to classification. Developing an approach to score features for interpretability remains an active area of research.

Interpretable Epigenetic Features

One area of work is to use prior knowledge of tissue- and cell-specific genome regulation to design interpretable features using non-coding genomic measures. We created a classifier that predicted binarized survival in Posterior Fossa A (PFA) ependymomas using patient DNA methylomes. PFA ependymomas are a subtype of a common pediatric neuroepithelial malignant tumour with an uncharacterized epigenetic component. Identifying cell types and epigenetic processes that predict prognosis in this cancer may lead to the development of actionable molecular therapies. Using processed DNA methylomes from Pajtler et al. [2018], we used netDx to classify patients as having good or poor prognosis (N=569 tumours, Illumina 450K microarrays). We compared performance of a basic model treating the entire methylome as single feature, to a "regulation-aware" design where base-level methylation was grouped into sets reflecting marker genes for individual cell types in the developing human cerebellum, binding sites for epigenetic regulators, and chromatin states

Submitted to the Machine Learning in Computational Biology 2021.

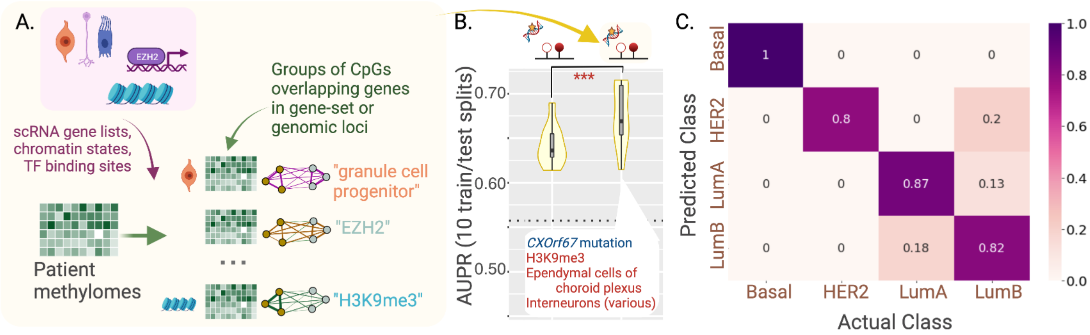

Figure 1: A. **Interpretable epigenetic features**: Conceptual schematic. B. Performance for binary survival prediction in PFA Ependymoma using DNA methylomes without (left) and with feature design reflecting brain genome regulation (right) (N=569; mean of 10 splits, p-value from one-sided WMW test). Pullout shows consistently high-scoring features. C. **Graph convolutional network-based classifier:** Confusion matrix for 4-way breast tumour classification by integrating gene expression and DNA methylation (N=511 tumours total; 52 in test set; 2 input networks).

in astrocytes and neural stem cells (Aldinger et al. [2021], ENCODE Project Consortium et al. [2020])(33 input networks). Samples were split 80:20 into train and test partitions, and training samples were used to score features out of 10. Features scoring 8 or higher were used to classify test samples. This process was repeated over 10 random train/test splits and model performance was measured (Fig. 1B). Features were defined as being consistently predictive if they scored 8 or higher in >= 70% of the train/test splits. We found that the predictor design aware of tissue-specific regulation significantly outperformed the model without this prior knowledge (Fig. 3B, median AUPR=0.67 for regulation-aware, AUPR=0.64 for other; $p < 4 \times 10^{-3}$, one-sided WMW test). Features passing selection capture affected cell types and the nature of chromatin dysregulation in PFA ependymoma, including ependymal cells from the developing human cerebellum and repressive chromatin state (H3K9me3 methylation) (Michealraj et al. [2020]). They also identify cell types not previously implicated in ependymoma, including interneurons of the molecular layer and unipolar brush cells, and excitatory cerebellar interneurons. Future work involves extending this strategy to other brain tumours to identify general principles in feature design for the non-coding genome.

Extension to deep learning

For improved scalability and discriminability, we are developing a classifier algorithm by extending a recently described approach for integrating multiple similarity networks using graph convolutional networks (GCN) (Forster et al. [2021]). BIONIC first encodes each user-input network separately using a GCN, then integrates these learned features. The integrated features can then be used for downstream tasks such as classification or clustering. To optimize its weights, BIONIC maps the integrated features back to the original input network adjacency matrices and minimizes the difference between them in an unsupervised manner. We converted this unsupervised algorithm to a classifier by adding a second cross-entropy term to the existing loss function and by providing the resulting embedding to a classifier, such as a support vector machine. Using this system, we classified breast tumours into one of four molecular subtypes by integrating gene expression and DNA methylation data (N=511 patients, 90:10 train/test split, 5-fold cross-validation) (TCGA Network [2012]). The model demonstrated an average F1-score of 0.88 across all 4 classes on the test set (Fig. 1C, N=52 samples; test accuracy=0.88).

A major remaining challenge is to identify a strategy for feature scoring, which is the basis for interpretability in our model. Explainable AI approaches such as LIME and SHAP are computationally infeasible for predictor designs with thousands of input features (such as pathway-based design) (Lundberg and Lee [2017], Ribeiro et al. [2016]). Moreover, saliency maps are not immediately adaptable to our design of integrating across multiple GCNs for node classification. This problem remains an area of active research.

References

B Wang, A M Mezlini, F Demir, M Fiume, Z Tu, M Brudno, B Haibe-Kains, and A Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, 11(3): 333–337, 2014.

Shraddha Pai and Gary D. Bader. Patient similarity networks for precision medicine. *Journal of Molecular Biology*, 430(18, Part A):2924–2938, 2018. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2018.05.037. URL `https://www.sciencedirect.com/science/article/pii/S0022283618305321`. Theory and Application of Network Biology Toward Precision Medicine.

Shraddha Pai, Shirley Hui, Ruth Isserlin, Muhammad A Shah, Hussam Kaka, and Gary D Bader. netdx: interpretable patient classification using integrated patient similarity networks. *Molecular Systems Biology*, 15(3):e8497, 2019. doi: https://doi.org/10.15252/msb.20188497. URL `https://www.embopress.org/doi/abs/10.15252/msb.20188497`.

Duncan T. Forster, Charles Boone, Gary D. Bader, and Bo Wang. Bionic: Biological network integration using convolutions. *bioRxiv*, 2021. doi: 10.1101/2021.03.15.435515. URL `https://www.biorxiv.org/content/early/2021/03/16/2021.03.15.435515`.

Kristian W Pajtler, Ji Wen, Martin Sill, Tong Lin, Wilda Orisme, Bo Tang, Jens-Martin Hübner, Vijay Ramaswamy, Sujuan Jia, James D Dalton, Kelly Haupfear, Hazel A Rogers, Chandanamali Punchihewa, Ryan Lee, John Easton, Gang Wu, Timothy A Ritzmann, Rebecca Chapman, Lukas Chavez, Fredrick A Boop, Paul Klimo, Noah D Sabin, Robert Ogg, Stephen C Mack, Brian D Freibaum, Hong Joo Kim, Hendrik Witt, David T W Jones, Baohan Vo, Amar Gajjar, Stan Pounds, Arzu Onar-Thomas, Martine F Roussel, Jinghui Zhang, J Paul Taylor, Thomas E Merchant, Richard Grundy, Ruth G Tatevossian, Michael D Taylor, Stefan M Pfister, Andrey Korshunov, Marcel Kool, and David W Ellison. Molecular heterogeneity and CXorf67 alterations in posterior fossa group a (PFA) ependymomas. *Acta Neuropathol.*, 136(2):211–226, August 2018.

Kimberly A Aldinger, Zachary Thomson, Ian G Phelps, Parthiv Haldipur, Mei Deng, Andrew E Timms, Matthew Hirano, Gabriel Santpere, Charles Roco, Alexander B Rosenberg, Belen Lorente-Galdos, Forrest O Gulden, Diana O'Day, Lynne M Overman, Steven N Lisgo, Paula Alexandre, Nenad Sestan, Dan Doherty, William B Dobyns, Georg Seelig, Ian A Glass, and Kathleen J Millen. Spatial and cell type transcriptional landscape of human cerebellar development. *Nat. Neurosci.*, June 2021.

ENCODE Project Consortium, Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L Van Nostrand, Peter Freese, David U Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A Williams, Ali Mortazavi, Cheryl A Keller, Xiao-Ou Zhang, Shaimae I Elhajjajy, Jack Huey, Diane E Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B Chhetri, Jialing Zhang, Alec Victorsen, Kevin P White, Axel Visel, Gene W Yeo, Christopher B Burge, Eric Lécuyer, David M Gilbert, Job Dekker, John Rinn, Eric M Mendenhall, Joseph R Ecker, Manolis Kellis, Robert J Klein, William S Noble, Anshul Kundaje, Roderic Guigó, Peggy J Farnham, J Michael Cherry, Richard M Myers, Bing Ren, Brenton R Graveley, Mark B Gerstein, Len A Pennacchio, Michael P Snyder, Bradley E Bernstein, Barbara Wold, Ross C Hardison, Thomas R Gingeras, John A Stamatoyannopoulos, and Zhiping Weng. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, July 2020.

Kulandaimanuvel Antony Michealraj, Sachin A Kumar, Leo J Y Kim, Florence M G Cavalli, David Przelicki, John B Wojcik, Alberto Delaidelli, Andrea Bajic, Olivier Saulnier, Graham MacLeod, Ravi N Vellanki, Maria C Vladoiu, Paul Guilhamon, Winnie Ong, John J Y Lee, Yanqing Jiang, Borja L Holgado, Alex Rasnitsyn, Ahmad A Malik, Ricky Tsai, Cory M Richman, Kyle Juraschka, Joonas Haapasalo, Evan Y Wang, Pasqualino De Antonellis, Hiromichi Suzuki, Hamza Farooq, Polina Balin, Kaitlin Kharas, Randy Van Ommeren, Olga Sirbu, Avesta Rastan, Stacey L Krumholtz, Michelle Ly, Moloud Ahmadi, Geneviève Deblois, Dilakshan Srikanthan, Betty Luu, James Loukides, Xiaochong Wu, Livia Garzia, Vijay Ramaswamy, Evgeny Kanshin, María Sánchez-Osuna, Ibrahim El-Hamamy, Fiona J Coutinho, Panagiotis Prinos, Sheila Singh, Laura K Donovan, Craig Daniels, Daniel Schramek, Mike Tyers, Samuel Weiss, Lincoln D Stein, Mathieu Lupien, Bradly G Wouters, Benjamin A Garcia, Cheryl H Arrowsmith, Poul H Sorensen, Stephane Angers, Nada Jabado, Peter B Dirks, Stephen C Mack, Sameer Agnihotri, Jeremy N Rich, and Michael D Taylor. Metabolic regulation of the epigenome drives lethal infantile ependymoma. *Cell*, 181(6):1329–1345.e24, June 2020.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.