Latent pathway modelling

Semester project

June, 2022

Project description

Biological pathways model series of (chronologically ordered) interactions among genes or proteins. Identification of these pathways is often a downstream goal of biological data analysis to summarize sets of identified active genes in a study and increase the interpretability of results (see Fig. 1). Most commonly, active pathways are identified via gene-set enrichment [GSEA, 6] or variants thereof via statistical testing. Since pathway analysis via GSEA is usually conducted after having identified a set of relevant genes in an upstream analysis, the results of the analysis depend on the previous steps of the statistical workflow and hence are prone to poor reproducibility.

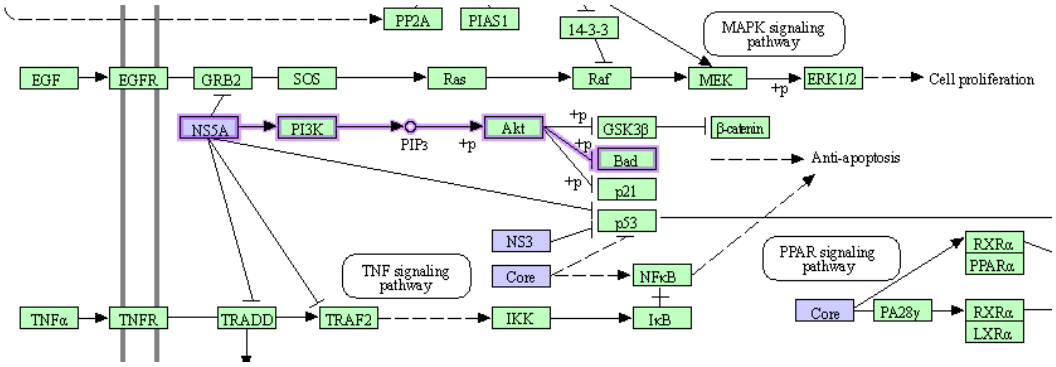

Figure 1: Excerpt of the Hepatitis C reference pathway (adapted from the *KEGG database* [4]). The viral non-structural protein 5A (shown in violet with frame) of Hepatitis C stimulates PI3K activity in the PIP pathway which downstream leads to reduction of apoptosis (induced cell death). Identification of the pathway itself rather than all genes involved is crucial for interpreting biological findings and clinical decision making.

Here, we propose to jointly infer the activity of genes and pathways via a Bayesian hierarchical factor model building on the work of Dirmeier and Beerenwinkel [2]. On the highest level of the hierarchy, the model quantifies the activity of a pathway and how much it activates (loads) a gene while the lower levels represent the observation model and its parameterization. Concretely, one could quantify the activity π_p of a pathway p as a variable on the unit interval which activates a gene g linearly with weight $w_{g,p}$:

$$\begin{aligned} \pi_p &\sim \text{Unif}\,(0,1) \\ w_{g,p} &\sim \text{Normal}(0,1) \;\; \text{if gene } g \text{ is part of pathway } p \\ \gamma_g \mid \mathbf{W}, \boldsymbol{\pi} &\sim \mathcal{N}\left(\mathbf{W}_{g,*}\boldsymbol{\pi}, \tau^2\right) \end{aligned} \tag{1}$$

The overall activity of a gene γ_g is then a linear combination of weights and all pathways that the gene is involved in. While this model is straightforward to implement, inference is complicated

because of, e.g., unidentifiability due to rotation invariance, for which we anticipate the model to require modern sparsity inducing priors [1, 5], incorporate additional constraints [3], etc. Having selected a competitive model, we will validate the model on a pan-viral genetic perturbation screen and compare it to conventional methods to identify pathways. The model will be a first step towards reproducible pathway analysis in a fully Bayesian framework and the simplification of workflows in biological data analysis. If the project is successful, the method should be published in appropriate journals or conferences.

Additional Information

- **What will you learn?**
  - Good understanding of Bayesian statistics, Markov chain Monte Carlo methods, structured sparsity
  - Genomics, virology, maybe cancer evolution
- **Requirements**
  - Basics in probabilistic modelling (hierarchical models, factor models) helpful
  - Python experience (!)
  - Experience working with high-dimensional biological data
  - Curiosity for biology (!) and enthusiasm for working with difficult data
- **Supervisor** Dr Simon Dirmeier (simon dot dirmeier at sdsc dot ethz dot ch) and Prof Dr Fernando Perez Cruz

References

[1] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. "Bayesian linear regression with sparse priors". In: *The Annals of Statistics* 43.5 (2015), pp. 1986–2018.

[2] Simon Dirmeier and Niko Beerenwinkel. "Structured hierarchical models for probabilistic inference from perturbation screening data". In: *bioRxiv* (2019).

[3] Elena A Erosheva and S McKay Curtis. "Dealing with rotational invariance in Bayesian confirmatory factor analysis". In: *Department of Statistics, University of Washington, Seattle, Washington, USA* (2011).

[4] *KEGG database*. https://www.genome.jp/kegg-bin/show_pathway?map=05160&orgs=hsa+pathogen. Accessed: 2022-06-01. 2000.

[5] Juho Piironen and Aki Vehtari. "Sparsity information and regularization in the horseshoe and other shrinkage priors". In: *Electronic Journal of Statistics* 11.2 (2017), pp. 5018–5051.

[6] Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.