Executive Summary

Introduction

Inferring latent chemical species in biological systems is a crucial task for understanding the underlying mechanisms of biological processes. Specifically, gaining insights into the regulation of gene expression and the interactions between transcription factors and DNA can provide valuable information on the mechanisms of diseases, drug side effects, and therapeutic innovations.[1][2][3] This research focuses on the transcription factor p53, known for its roles in cancer suppression and cellular stress responses, utilising a Latent Force Model (LFM) to model the unobserved activity profile of p53 from the gene expressions of its target genes. [4][5]

Background

Barenco et al., 2006 proposed a biologically-motivated non-autonomous linear differential equation to model the rate of change in transcript concentration $x_j(t)$ of target gene j at time t given by Equation (1):

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j f(t) - D_j x_j(t) \tag{1.1}$$

The authors justified this model by arguing that the rate at which p53-dependent mRNA transcripts accumulate (measured as the rate of gene expression) depends on the base transcription rate of a target gene (B), the sensitivity of that gene to p53 (S), the rate of mRNA decay (D), and the activity level of p53 itself (f).[4]

Assuming an initial baseline expression level $x_j(0) = B_j/D_j$, Equation (1) can be solved to recover

$$x_j(t) = \frac{B_j}{D_j} + L_j[f](t) \tag{1.2}$$

Where $L_j[f](t)$ is the linear operator (convolution integral) relating the latent function f to the mRNA abundance of gene j, $x_j(t)$:

$$L_j[f](t) = S_j e^{-D_j t} \int_0^t e^{D_j u} f(u) \mathrm{d}u \tag{1.3}$$

Elementary functional analysis yields that:

$$\mathrm{Cov}(L_j[f](t), L_k[f](t')) = L_j \otimes L_k k_{ff}(t, t') \tag{1.4}$$

Lawrence et al., 2006 then treated the protein concentration of p53 as a latent function with a Gaussian process prior using the squared exponential kernel. Through Equations (1) and (1), the authors were able to obtain analytical solutions for the covariance function of gene expressions at times t' and t: $k_{x_j x_k}(t, t')$ as well as the cross-covariance between measured gene expression at time t and the latent force function at time t': $k_{x_j f}(t, t')$. They then used Gaussian process regression to infer the latent force function $f(t)$ from the gene expression data.[5]

Objectives & Scope

This project aims to replicate and extend the findings of Lawrence et al. (2006) by implementing the LFM in GPJax, a Gaussian process library built on JAX.[6] This involved building a custom model to handle the three kernels used in the LFM, as well as a custom trainer to fix specific hyperparameters of the model, making the results experimentally identifiable to those obtained by Barenco et al. (2006) and Lawrence et al. (2006). The project also aimed to investigate the model's performance on ablated data and different replicas of the same experiment.

Methodology

To implement this latent force model in GPJax, which does not natively support vector-valued functions, the dimensionality of the gene expression dataset was augmented to include the timepoints, gene indices, and an additional *flag* dimension so that the data points are three-dimensional: $d = (t, j, z)$ where t is the timepoint, j is the gene index, and $z \in \{0, 1\}$. Using this flag, a GP was used to learn the function $g(\bullet)$, which is defined as follows:

$$g(t, i, z) = \begin{cases} x(t, i) & \text{if } z = 1, \\ f(t) & \text{otherwise.} \end{cases} \tag{1.5}$$

where $x(\bullet)$ is the gene expression as a function of the gene index and timepoint, and $f(\bullet)$ is the latent function at time t.

For two inputs $\mathbf{T} = (t, j, z)$ and $\mathbf{T}' = (t', k, z')$, the combined kernel function $k(\mathbf{T}, \mathbf{T}')$ is defined as follows:

$$k(\mathbf{T}, \mathbf{T}') = \begin{cases} k_{xx}((t, j), (t', k)) & \text{if } z = z' = 1 \\ k_{xf}((t, j), t') & \text{if } z \neq z' \\ k_{ff}(t, t') & \text{if } z = z' = 0 \end{cases} \tag{1.6}$$

Where $k_{xx}((t, j), (t, k))$ and $k_{xf}((t, j), t')$ are modifications of $k_{x_j x_k}(t, t')$ and $k_{x_j f}(t, t')$, respectively, to account for the additional gene index dimension.

The LFM containing this custom kernel was trained with the sensitivity and decay rate of target gene p21 fixed at 1.0 and 0.8, respectively, following the implementation by Lawrence et al., to make the inferred hyperparameters identifiable.

Key findings

The latent activity profile of p53 predicted by the LFM implemented in GPJax for this project is shown in Figure 1.1 below. The solid line is the posterior mean prediction, dashed lines represent two standard deviations, and experimental values are shown as crosses.

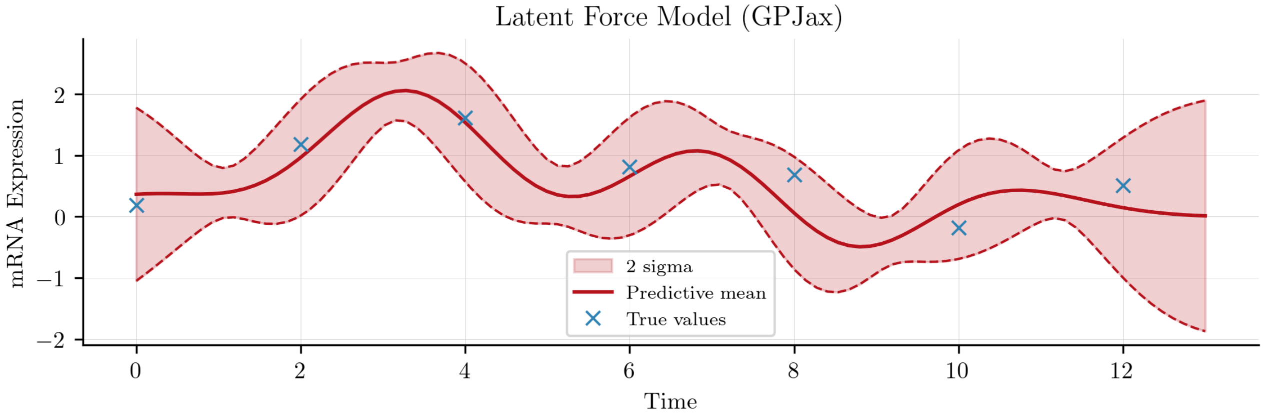

Figure 1.1: Predicted p53 activity profile from GPJax.

The posterior mean function aligns closely with the experimental predictions, especially in the region between time zero and time six, and is in good agreement with the results obtained by the authors of the original paper. Hyperparameters obtained from the optimised model are shown in Figure 1.2 below. The red bars are estimates obtained from the posterior distribution and the blue bars are experimental values obtained by Barenco et al., 2006.

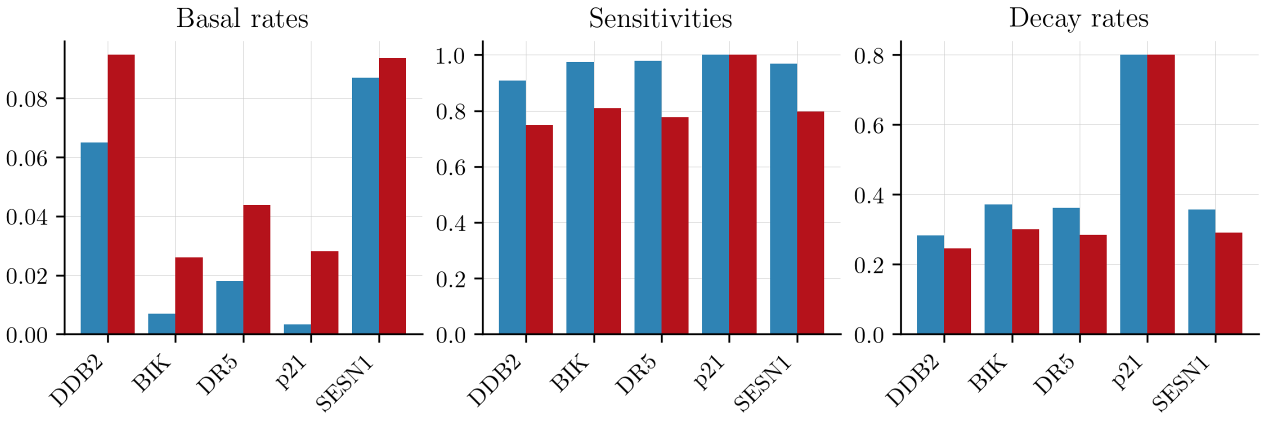

Figure 1.2: Inferred hyperparameters for p53 target genes from GPJax.

In addition to successfully reproducing the results obtained by Lawrence et al., 2006 and validating these results with a linear force model from the ALFI package, the robustness of the LFM was tested by performing an ablation study and training the model on different replicas of the same dataset.[7] Ablation study results showed that the model was able to accurately predict the latent activity of p53 when trained on as few as three target gene, challenging the need to train the model with all five target genes and suggesting that the initial dataset of five genes may not have provided substantial additional information. When trained on different gene replicas, the model's performance was inconsistent: the posterior distributions obtained from the second and third replicas deviating significantly from the experimental data.

Conclusions

In summary, it was possible to reproduce the key results from Lawrence et al., 2006. However, the lack of details regarding data preprocessing and implementation from the authors, coupled with the varied results obtained from training the model on ablated data and different replicas, complicated the verification of whether the selection of data in the original study was influenced by performance metrics.

Word count: 836

References

[1] S. Huang, "Gene expression profiling, genetic networks, and cellular states: An integrating concept for tumorigenesis and drug discovery," *Journal of Molecular Medicine*, vol. 77, no. 6, pp. 469–480, Jun. 1999, ISSN: 1432-1440. DOI: 10.1007/s001099900023. [Online]. Available: http://dx.doi.org/10.1007/s001099900023.

[2] P. A. Clarke, R. te Poele, R. Wooster, and P. Workman, "Gene expression microarray analysis in cancer biology, pharmacology, and drug development: Progress and potential," *Biochemical Pharmacology*, vol. 62, no. 10, pp. 1311–1336, 2001, ISSN: 0006-2952. DOI: https://doi.org/10.1016/S0006-2952(01)00785-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006295201007857.

[3] E. R. Gibney and C. M. Nolan, "Epigenetics and gene expression," *Heredity*, vol. 105, no. 1, pp. 4–13, May 2010, ISSN: 1365-2540. DOI: 10.1038/hdy.2010.54. [Online]. Available: http://dx.doi.org/10.1038/hdy.2010.54.

[4] M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank, *Genome Biology*, vol. 7, no. 3, R25, 2006, ISSN: 1465-6906. DOI: 10.1186/gb-2006-7-3-r25. [Online]. Available: http://dx.doi.org/10.1186/gb-2006-7-3-r25.

[5] N. Lawrence, G. Sanguinetti, and M. Rattray, "Modelling transcriptional regulation using gaussian processes," vol. 19, 2006. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2006/hash/f42c7f9c8aeab0fc412031e192e2119d-Abstract.html.

[6] T. Pinder and D. Dodd, "Gpjax: A gaussian process framework in jax," *Journal of Open Source Software*, vol. 7, no. 75, p. 4455, 2022. DOI: 10.21105/joss.04455. [Online]. Available: https://doi.org/10.21105/joss.04455.

[7] J. D. Moss, F. L. Opolka, B. Dumitrascu, and P. Lió, *Approximate latent force model inference*, Jan. 24, 2022. DOI: 10.48550/arXiv.2109.11851. arXiv: 2109.11851[cs]. [Online]. Available: http://arxiv.org/abs/2109.11851.