GPCell: A Performant Framework for Gaussian Processes in Bioinformatics

Tristan Sones-Dykes

MMath Mathematics

April 2025

School of Mathematics and Statistics

The University of St Andrews

Submitted in total fulfilment of the requirements of the degree of MMath Mathematics

Abstract

Gene expression regulation is pivotal in cellular function, with significant advancements since the 1960s. Notably, Jacob and Monod's work elucidated gene activation mechanisms in response to external stimuli via mRNA transcription modulation (Jacob & Monod, 1961). Subsequent research, such as Hardin *et al.* (1990)'s study on circadian rhythms in *Drosophila melanogaster*, highlighted oscillatory gene expression through protein-mediated RNA inhibition. Building upon these foundations, Phillips *et al.* (2017) investigated gene expression patterns in neural progenitor cells, identifying correlations between oscillatory behavior and differentiation. Their methodology employed Gaussian processes to classify gene expression time series using MATLAB.

This dissertation extends their approach by developing a Python library that facilitates Gaussian process fitting and oscillation detection across diverse datasets. Enhancements include an extensible modelling framework, allowing for the easy addition of fitting techniques like MCMC; being based coherently on top of Tensorflow Probability, taking advantage of computational advancements and giving access to a suite of priors, model types, and optimisers; and an automated Continuous Integration/Continuous Deployment (CI/CD) pipeline, with accuracy tests for models and automatically generated docs (Sones-dykes, 2025).
Future works can now prioritise scientific discovery and model choice, using a suite of utilities that simplify the model-fitting process.

Table of contents

1 Introduction

The regulation of gene expression is fundamental to cellular processes, dictating how genes are activated or repressed in response to various stimuli. A landmark discovery by Jacob and Monod in the 1960s (Jacob & Monod, 1961) demonstrated that genes could be switched on or off through the modulation of mRNA transcription, a mechanism that garnered them the Nobel Prize. While their research provided insights into gene regulation mechanisms, it did not delve into the temporal dynamics of gene expression.

In the 1980s, research into the temporal aspects of gene expression gained momentum. Hardin *et al.* (1990) proposed the Transcription Translation Negative Feedback Loop to explain circadian rhythms in *Drosophila melanogaster*. They discovered that certain genes exhibited oscillatory behavior, regulated by proteins that inhibited their own RNA production, thereby establishing a feedback loop.

Furthering this line of inquiry, Phillips *et al.* (2017) explored gene expression patterns in neural progenitor cells, revealing that such expression could be either oscillatory or aperiodic. Notably, oscillatory expression was associated with cellular differentiation. Their methodology involved fitting both oscillating and non-oscillating Gaussian processes to gene expression time series data, enabling the classification of genes based on their oscillatory behavior. However, their approach was primarily applied within a specific biological context, limiting its generalisability; it was also developed in MATLAB, narrowing its possible scope and extensibility.

This dissertation aims to broaden the applicability of Phillips *et al.* (2017)'s methodology by developing a testable and extensible Python library. This library is designed to facilitate model-fitting and oscillation detection across various datasets and experimental conditions. Utilizing the GPflow library, which is based on TensorFlow, it provides a user-friendly interface for researchers, with an OscillatorDetector class specifically designed for this use-case. It also has unit tests to verify correctness, a strong type system to aid development and upkeep, as well as a suite of abstracted utility functions that automate the model fitting process through a multiprocessing pipeline increasing fitting speed.

2 Methods

3 Results

4 Discussion

References

Hardin, P.E., Hall, J.C., & Rosbash, M. (1990) Feedback of the Drosophila period gene product on circadian cycling of its messenger RNA levels. *Nature*, 343, 536–540.

Jacob, F. & Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3, 318–356.

Phillips, N.E., Manning, C., Papalopulu, N., & Rattray, M. (2017) Identifying stochastic oscillations in single-cell live imaging time series using Gaussian processes. *PLOS Computational Biology*, 13, e1005479.

Sones-dykes, T. (2025) GPCell documentation.

List of figures

List of tables