GPCell: A Performant Framework for Gaussian Processes in Bioinformatics

Tristan Sones-Dykes

School of Mathematics and Statistics, The University of St Andrews

2025-04-01

Agenda

- Motivation & Background
- Problem Statement
- GPCell: Your Solution
- Methods & Results Overview
- Discussion & Impact
- Future Work
- Conclusion & Takeaways
- Assessment Criteria
- Q & A

Motivation & Background

- **Importance of Gene Expression Regulation**
  - Fundamental to cellular processes
- **Historical Foundations**
  - Jacob and Monod (1961)
  - Hardin, Hall, and Rosbash (1990)
  - Phillips et al. (2017)
- **Introduction to Gaussian Processes**
  - Non-parametric, probabilistic, and flexible modeling approach

Problem Statement

- **Limitations in Prior Work:**
  - MATLAB-based approach in Phillips et al. (2017)
  - Domain-specific application and limited extensibility
- **The Need for GPCell:**
  - A general, extensible Python framework for GP fitting and oscillation detection

GPCell: Your Solution

- **Overview:**
  - A Python library built on TensorFlow Probability
- **Key Features:**
  - **OscillatorDetector:** Class for analyzing gene expression oscillations
  - Extensible framework (e.g., integration of MCMC methods)
  - Strong type system and automated CI/CD pipeline
  - Multiprocessing pipeline for improved performance
- *(Consider adding code snippets or diagrams here)*

Methods & Results Overview

- **Validation:**
  - Tested on diverse datasets (synthetic and real data)
- **Performance:**
  - Speed improvements using multiprocessing
  - Accuracy in oscillation detection
- **Visuals:**
  - Include charts/diagrams to illustrate model fitting and results

Discussion & Impact

- **Strengths:**
  - Extensibility and user-friendly design
  - Enhanced reproducibility for bioinformatics research
- **Challenges:**
  - Managing performance bottlenecks
  - Addressing edge cases in complex datasets
- **Broader Impact:**
  - Provides a robust platform for advanced GP modeling in bioinformatics

Future Work

- **Enhancements:**
  - Integration of additional kernels and priors
  - Further development of visualization tools
- **Applications:**
  - Integration with wet lab pipelines for real-time data analysis

Conclusion & Takeaways

- GPCell bridges the gap between advanced statistical methods and bioinformatics.
- It provides a scalable, extensible, and reproducible framework for Gaussian Process modeling.
- Opens new avenues for scientific discovery and research efficiency.

Assessment Criteria

- **Content:**
  - Substance, structure, and depth of understanding
  - Accuracy of the presented information
- **Delivery:**
  - Clarity of communication and engagement
  - Effective use of slides and visual aids
- **Overall Presentation Quality:**
  - Professionalism and impact of the presentation

Q & A

Thank you!
Questions?

Hardin, Paul E., Jeffrey C. Hall, and Michael Rosbash. 1990. "Feedback of the Drosophila Period Gene Product on Circadian Cycling of Its Messenger RNA Levels." *Nature* 343 (6258): 536–40. https://doi.org/10.1038/343536a0.

Jacob, François, and Jacques Monod. 1961. "Genetic Regulatory Mechanisms in the Synthesis of Proteins." *Journal of Molecular Biology* 3 (3): 318–56. https://doi.org/10.1016/S0022-2836(61)80072-7.

Phillips, Nick E., Cerys Manning, Nancy Papalopulu, and Magnus Rattray. 2017. "Identifying Stochastic Oscillations in Single-Cell Live Imaging Time Series Using Gaussian Processes." *PLOS Computational Biology* 13 (5): e1005479. https://doi.org/10.1371/journal.pcbi.1005479.