Deep Learning methods in Genomic Medicine

Matthew Ramcharan
Supervised by Dr Colin Campbell

November 29, 2018

1 Introduction

Somatic mutations are any alteration in cell that will not be passed onto future generations[2]. A somatic mutation in a cell of a fully developed organism can have little to no noticeable effect on the organism itself (often leading to benign growths), however mutations that give rise to cancer are a special case. Cancer mutations arise in a special category of genes called proto-oncogenes, many of which regulate cell division. When mutated, such cells enter a state of uncontrolled division and become Oncogenes, resulting in a cluster of cells called a tumor. There is also mutations in tumor-suppressor genes which can cause tumors by preventing regulation of cell proliferation.

These types of cell division lead to malignant tumors, in which the excessive cell proliferation causes the tumor to spread into surrounding tissues and cause damage.

This means being capable of discriminating between benign and oncogenic mutations is integral to identifying cancer before the tumor gets too large, or in grows to be in a bad position to excise. In this project we implement each

2 Literature review

CScape was cool, lets make it better. Gunnar Rasch has a great idea for genomes.

3 Project plan

This project is primarily following this plan:

The models that will be built with these datasets are

1. Support Vector Machine - The simplest kernel method for integrating different data sources

is to combine the features from all sources into a single kernel [3]. Creates a binary classifier of either Pathogenic (Positive) or Control (Negative)

2. Multiple Kernel Learning

3. Multi-task Multiple Kernel Learning

Action	Timeframe	Project Relevance
Research the problem	Weeks 2-4	The dataset style used in the CScape [3] and the method proposed in the Framework for Multi-task Multiple Kernel Learning [4] are integral to this project, and so should be understood well.
Learn how to apply the Shogun toolkit for toy problems	Weeks 4-5	Learning how to use an existing implementation of the Multi-task Multiple Kernel Learning will provide greater insight into expected outputs when used on more complex datasets.
Download and understand CScape dataset	Week 5	Simplest form of dataset to be used in this project
Gain an understanding of the COSMIC [1] data	Weeks 5-6	Selecting data to be used from the vast database COSMIC will be what is used for the main stage of this project - the implementation of Multi-task, multiple kernel learning.
Understand existing models (CScape, FATHMM-MKL,	Weeks 6-9	
Write Interim Report	Weeks 9-10	
	Weeks 10-12	.
	Weeks 12	
	Christmas!	
	Week 13	
Ready for, and present Presentation	Week 14	
	Weeks 14-16	
Finish Technical work, draw conclusions and consolidate report	Weeks 16-19	Allow plenty of time to finish and polish the report.
Draft of one chapter/section of final report hand-in.	Week 17	Section is written during consolidation period

Update/change report style based on feedback of single section and continue writing	Week 17-19	Report changes are still during consolidation period
Create Poster	Week 19	
Proofread Report	Week 19-20	
Submit full draft of final report and poster	Week 20	
Update report and poster with any relevant details from Supervisor Dr Colin Campbell	Week 21	
Proofread	Week 22	
Final hand-in date for report and poster	Week 22.3 (First week of easter)	Noted for completeness, hand-in should be late Week 21/ early Week 22

4 Progress

References

[1] Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymund Stefancsik, Bhavana Harsha, Chai Yin Kok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam Thompson, Tisham De, and Peter J. Campbell. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, jan 2017.

[2] AJF Griffiths, JH Miller, and DT Suzuki. An Introduction to Genetic Analysis. 7th edition. chapter 15. New York: W. H. Freeman, 7th edition, 2000.

[3] Mark F. Rogers, Hashem A. Shihab, Tom R. Gaunt, and Colin Campbell. CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Scientific Reports*, 7(1):11597, dec 2017.

[4] Christian Widmer, Marius Kloft, Vipin T Sreedharan, and Gunnar Rätsch. Framework for Multi-task Multiple Kernel Learning and Applications in Genome Analysis. jun 2015.