Deep Learning methods in Genomic Medicine
Interim Report

Matthew Ramcharan
Supervised by Dr Colin Campbell

December 4, 2018

1 Introduction

Somatic mutations are any alteration in cell that will not be passed onto future generations [3]. 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 arises either from inactivation of tumor suppressor genes, or mutation of a special category of genes called proto-oncogenes, many of which regulate cell division. When mutated, proto-oncogenes enter a state of uncontrolled division and become Oncogenes, resulting in a cluster of cells called a tumor. 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.

Genome sequences are stored as Singular nucleotide polymorpisms, which are the difference in a single DNA building block, called a nucleotide. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene's function.

In this project we focus on the prediction for somatic point mutation in the coding region of the human cancer genome. There are many cancer sequence databases currently being compiled, such as the Cancer Genome Atlas, COSMIC,

We will call the new method when used on general data tentatively FATHMM-MTMLK CScape-MTMLK

the portion of the genome which codes for proteins accounts for only about 2% of the whole sequence, and it is becoming increasingly evident that non–coding portions of the genome play crucial functional roles in human development and disease[1]

2 Literature review

The problem of identifying which variations in genomic information drive disease is a well recorded and explored one

2.1 General Purpose Predictors

general-purpose pathogenic mutation classifiers across coding and noncoding regions. [5, 4]

2.2 Cancer Specific Predictors

CScape was cool, lets make it better.

[6, 5, 7].

The specific case of if the disease driven is cancer Rogers et al [6]

2.3 Multi-task Multiple Kernel Learning

Gunnar Rasch has a great idea for genomes. [9]

3 Project plan

This project is primarily following this plan:

The datasets used

1. CScape data [6] - Contains labelled genome data for somatic and germline SNVs which are distingushed
2. COSMIC Data -

The algorithms that will be used to predict 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 [6]. Creates a binary classifier of either Pathogenic (Positive) or Control (Negative)
2. Multiple Kernel Learning - A composite kernel is made from a set of base kernels, in which each base kernel is derived from an individual set of data, which are different features like Histone Modifications, 100-Way Sequence Conservation, or Genome Segmentation. [8]

3. Multi-task Multiple Kernel Learning - Applying the multiple kernel with multi-tasks where each task is a type of cancer (lung, breast, brain, etc) [9]

Results that need to be produced

1. ROC curve of

Action	Timeframe	Project Relevance
Research the problem	Weeks 2-4	The dataset style used in the CScape [6] and the method proposed in the Framework for Multi-task Multiple Kernel Learning [9] 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 [2] 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, 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

5 Conclusions and Further Work

References

[1] Manel Esteller. Non-coding RNAs in human disease. *Nature Reviews Genetics*, 12(12):861–874, dec 2011.

[2] 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.

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

[4] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, mar 2014.

[5] Daniel Quang, Yifei Chen, and Xiaohui Xie. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, mar 2015.

[6] 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.

[7] Hashem A. Shihab, Julian Gough, David N. Cooper, Ian N. M. Day, and Tom R. Gaunt. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*, 29(12):1504–1510, jun 2013.

[8] Hashem A. Shihab, Mark F. Rogers, Julian Gough, Matthew Mort, David N. Cooper, Ian N. M. Day, Tom R. Gaunt, and Colin Campbell. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, 31(10):1536–1543, may 2015.

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