Deep Learning methods in Genomic Medicine
Interim Report

Matthew Ramcharan
Supervised by Dr Colin Campbell

December 5, 2018

1 Introduction

Somatic mutations are any alteration in cell that will not be passed onto future generations [7]. 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 tumour 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 tumour. These types of cell division lead to malignant tumours, in which the excessive cell proliferation causes the tumour to spread into surrounding tissues and cause damage.

A common, probably simplistic, model view defines two classes of mutations, 'driver' mutations, i.e. mutations that give a cancer cell a particular selective advantage, and functionally irrelevant 'passenger' mutations. Discovering functionally important mutations, including clear 'drivers' is one goal of genome re-sequencing efforts [13]. To understand the functional contribution of molecular alterations to oncogenesis, response to therapy and evolution of resistance to therapy it is important to have tools that predict the functional implications of mutations as early in the discovery process as possible.

As a dataset, genome sequences are stored as Singular nucleotide polymorphisms, which are the difference in a single DNA building block, called a nucleotide. When SNPs occur within a gene (the coding region) or in a regulatory region near a gene (certain parts of the non-coding regions), they may play a more direct role in disease by affecting the gene's function.

The coding region, 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[4]. This implies there is merit to attempting the same methods on data from both the coding and non-coding regions of the human genome.

In this project we focus on prediction of the effects of somatic point mutations leading to amino acid substitutions[15] in the coding and noncoding region of the human cancer genome. These predictions will be assigned a label as to if a point mutation is oncogenic (Likely cancerous) or benign. There are many cancer sequence databases currently being compiled, such as the Cancer Genome Atlas, COSMIC, and the National Cancer Institute and an large aspect of this project is selecting appropriate data to correctly train a cancer predictor, then test it holds up to a variety of data sources.

2 Literature review

The problem of identifying which variations in genomic information drive disease is a well recorded and explored one. This project will focus on developing a cancer specific predictor, however the sequencing techniques in identifying any deleterious mutations are similar for most diseases. Cancer specific predictors are still advancing every day, to the point they are starting to become useful in clinical applications[3].

2.1 General Purpose Predictors

General-purpose pathogenic mutation classifiers across coding and non-coding regions have been implemented using multiple different machine learning architectures, including deep learning[12], support vector machines [8] and Multiple Kernel Learning [16]. Since there are so many pathogenic variants recorded, and their effects are better known, these models can be considered somewhat easier than cancer specific predictors to train and test.

2.2 Cancer Specific Predictors

[14, 12, 15].

The specific case of if the disease driven is cancer Rogers et al [14]

2.3 Multi-task Multiple Kernel Learning

Widmer et al [17] proposed an advancement on Multi-task learning which utilises the results from Multiple Kernel learning to inform the similarities between tasks.

3 Project plan

This project is primarily following this plan:

3.1 Datasets

The datasets that will be used will be

1. CScape data [14] - Contains labelled genome data for somatic and germline SNVs which can be distinguished into different types of cancer. Uses data from COSMIC and the 1,000 Genomes Project [6]
2. COSMIC Data - Shown to be a good source for cancer somatic point mutations. Supplementary Table SM2 of Reva et al[13] shows 957 genes with significance for cancer, of which there are many unique mutations for each
3. ClinVar[9] - As a set of unseen test examples.

3.2 Models

The models that will be used to predict labels for 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 [14]. 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 [16].
3. Multi-task Multiple Kernel Learning[17] - Applying the multiple kernel learning method with multi-tasks where each task is a type of cancer (lung, breast, brain, etc.)

3.3 Results

Results and measures that the models will output are:

1. Test (and training) accuracies of each method on a given training and test set using Leave One Chromosome Out Cross Validation (LOCO-CV) [14, 18, 10, 11] in which for each fold we leave out one test chromosome while the remaining 21 chromosomes are used to train the model, using the same model parameters for all folds (similar to 22-fold cross-validation). to achieve a high balanced accuracy[2] to derive an accurate estimate of label prediction performance.
2. ROC curves for every method tried on a dataset along with their AUCs.

3.4 Action Plan

Action	Time frame	Project Relevance
Research the problem	Weeks 2-4	The dataset style used in the CScape [14] and the method proposed in the Framework for Multi-task Multiple Kernel Learning [17] 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	This is, in a way, the simplest form of dataset to be used in this project.
Gain an understanding of the COSMIC [5] 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	Having a high level understanding of the existing models will help to know how the problem has been approached, and if there are any unusual processes when implementing a kernel based method with genomic data.
Write Interim Report	Weeks 9-10	
Implement Simple Support Vector Machine on CScape data	Week 10	Working up from the simplest baseline in the hierarchy of models will allow greater insight into how the models function, and show how the methods vary in accuracy.
Implement Multiple Kernel Learning on CScape data	Week 11	Similar to previous week.
Collect CScape equivalent datasets from COSMIC and 1000 Genome project including cancer type	Week 12	The key difference between the existing CScape dataset and the one necessary for MTMKL is that each genome sequence should have the type of cancer the sample came from labelled.
Implement Multi-task Multiple Kernel Learning on CScape data	Weeks 12-13	Longer term implementation. This forms the basis of a novel approach to problem. This will allow us to generate results to compare this novel method to the previous ones.
	Christmas Break	

Generate results for different models. Primarily ROC curves and accuracy measurements	Week 13	These results should be the primary discussion point for the MTMKL model, distinguishing it from existing models.
Ready for, and present Presentation	Week 14	
Collect data from unseen data sources (as discussed above) and preprocess to be appropriate to all implemented models	Week 15-16	Testing the models on previously unseen data sources will reveal if there is any bias in the CScape dataset.
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 the Week 16-19 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 Conclusions and Further Work

Most modern advancement in the relevant general purpose nonsynonymous nucleotide variant predictor field is Alirezaie et al[1] but is currently previous planned work will need to be completed before the decision to invest time and money into this paper can be made.

References

[1] Najmeh Alirezaie, Kristin D. Kernohan, Taila Hartley, Jacek Majewski, and Toby Dylan Hocking. "ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants". In: *The American Journal of Human Genetics* 103.4 (Oct. 2018), pp. 474–483. ISSN: 00029297. DOI: 10.1016/j.ajhg.2018.08.005.

[2] Kay H Brodersen, Cheng Soon Ong, Klaas E Stephan, and Joachim M Buhmann. "The balanced accuracy and its posterior distribution". In: (2010). DOI: 10.1109/ICPR.2010.764.

[3] Chiara Di Resta and Maurizio Ferrari. "Next Generation Sequencing: From Research Area to Clinical Practice." In: *EJIFCC* 29.3 (Nov. 2018), pp. 215–220. ISSN: 1650-3414.

[4] Manel Esteller. "Non-coding RNAs in human disease". In: *Nature Reviews Genetics* 12.12 (Dec. 2011), pp. 861–874. ISSN: 1471-0056. DOI: 10.1038/nrg3074.

[5] Simon A. Forbes et al. "COSMIC: somatic cancer genetics at high-resolution". In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D777–D783. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1121.

[6] Richard A. Gibbs et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393.

[7] AJF Griffiths, JH Miller, and DT Suzuki. "An Introduction to Genetic Analysis. 7th edition." In: 7th. New York: W. H. Freeman, 2000. Chap. 15. ISBN: 0-7167-3520-2.

[8] 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". In: *Nature Genetics* 46.3 (Mar. 2014), pp. 310–315. ISSN: 1061-4036. DOI: 10.1038/ng.2892.

[9] Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. "ClinVar: public archive of relationships among sequence variation and human phenotype". In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D980–D985. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1113.

[10] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. "FaST linear mixed models for genome-wide association studies". In: *Nature Methods* 8.10 (Oct. 2011), pp. 833–835. ISSN: 1548-7091. DOI: 10.1038/nmeth.1681.

[11] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. "Improved linear mixed models for genome-wide association studies." In: *Nature methods* 9.6 (May 2012), pp. 525–6. ISSN: 1548-7105. DOI: 10.1038/nmeth.2037.

[12] Daniel Quang, Yifei Chen, and Xiaohui Xie. "DANN: a deep learning approach for annotating the pathogenicity of genetic variants". In: *Bioinformatics* 31.5 (Mar. 2015), pp. 761–763. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu703.

[13] Boris Reva, Yevgeniy Antipin, and Chris Sander. "Predicting the functional impact of protein mutations: application to cancer genomics." In: *Nucleic acids research* 39.17 (Sept. 2011), e118. ISSN: 1362-4962. DOI: 10.1093/nar/gkr407.

[14] 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". In: *Scientific Reports* 7.1 (Dec. 2017), p. 11597. ISSN: 2045-2322. DOI: 10.1038/s41598-017-11746-4.

[15] 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". In: *Bioinformatics* 29.12 (June 2013), pp. 1504–1510. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btt182.

[16] Hashem A. Shihab et al. "An integrative approach to predicting the functional effects of non-coding and coding sequence variation". In: *Bioinformatics* 31.10 (May 2015), pp. 1536–1543. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btv009.

[17] Christian Widmer, Marius Kloft, Vipin T Sreedharan, and Gunnar Rätsch. "Framework for Multi-task Multiple Kernel Learning and Applications in Genome Analysis". In: (June 2015). arXiv: 1506.09153.

[18] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. "Advantages and pitfalls in the application of mixed-model association methods." In: *Nature genetics* 46.2 (Feb. 2014), pp. 100–6. ISSN: 1546-1718. DOI: 10.1038/ng.2876.