Cancer Progression Classification for Mutation Analysis

Jari Björne and Tapio Salakoski

Department of Information Technology, University of Turku
Turku Centre for Computer Science (TUCS)
Joukahaisenkatu 3-5, 20520 Turku, Finland
firstname.lastname@utu.fi

1 Introduction

The International Cancer Genome Consortium is a global project aiming to produce a description of the genomic, transcriptomic and epigenomic changes in 50 different cancer types. The current 18th release of the ICGC data provides varied biochemical analyses for a set of 12,807 cancer patients. The CAMDA conference concerns the development of computational approaches for analysis of large-scale biomedical datasets. The ICGC dataset has been used in the shared "CAMDA challenge" first in 2014 and now in 2015. Having participated in the 2014 challenge we now extend our work of analysing cancer datasets through functional classification followed by analysis of the genetic basis behind the classification. In our 2014 entry we studied the ICGC cancers as separate datasets and evaluated various approaches for feature selection [Björne et al., 2014]. In the current work we expand our analysis to the whole ICGC cross-cancer dataset. With the increased scale of the data, we are able to utilize the somatic mutation data that illuminates the underlying causes of the cancer. With the effective embedded feature analysis of an ensemble classifier we can evaluate the type of mutations that affect the clinical outcome of a patient's cancer. Comparison with the COSMIC cancer gene census shows that genes central in causing cancer are also central for predicting its progression.

2 Materials and Methods

2.1 The ICGC cancer data

We use the publicly available parts of release 18 of the ICGC project. Files for the 55 cancer projects were downloaded from the ICGC data portal[1] [Zhang et al., 2011]. For use in experiments the TSV-formatted files were converted into an SQLite database approximately 30 Gb in size.

2.2 COSMIC cancer gene census

The COSMIC database (Catalogue Of Somatic Mutations In Cancer) is a collection of somatic mutations present in cancers, developed by the Wellcome Trust Sanger Institute[2]. The COSMIC cancer gene census is a list of genes known to be causally implicated in cancer [Futreal et al., 2004]. It thus represents a conservative set of the most strongly cancer related genes. The version of the census used in these experiments was downloaded on May 18th 2015 and consists of 572 genes.

2.3 Machine learning methods

For machine learning we use the scikit-learn library, version 0.16.1. We perform binary classification, using the Linear SVC (support vector machine) and Extra Trees classifiers [Geurts et al., 2006]. In the current 0.16.1 release of scikit-learn both of these methods support sparse matrices allowing efficient processing of large data sets.

[1] https://dcc.icgc.org
[2] http://www.sanger.ac.uk/cosmic

For estimating classification performance we use the scikit-learn implementation of the AUC-metric (area under the ROC curve). The AUC is a robust and largely class-distribution independent performance measure, whose results are in the range 0.5 (completely random) to 1.0 (perfect classification).

3 Experimental Setup

3.1 Division of Data

In performing classification experiments we optimize parameters using five-fold cross-validation on a *training* dataset. Final results are produced on a separate *hidden* dataset left aside for this purpose. We divide ICGC cancer samples by patient into training and hidden sets in a 7:3 ratio. The sets are divided on a pseudorandom distribution seeded with the *ICGC donor id*, ensuring that the same patient always belongs to either the training or the hidden set regardless of the selection of patients for a particular experiment.

3.2 Classification

Our goal is to develop a classification system for predicting the prognosis of a patient's cancer based on the available biochemical data. The prognosis is of interest as a classification task in itself, but also as a preliminary step for the feature analysis that aims to uncover the genetic basis of the prognosis.

The primary classification we perform is the division of the ICGC cancers that go into "complete remission" (disappearance of all signs of cancer) vs. those that progress to the death of the patient. These represent the two, opposite end-points for a cancer patient. This division follows our per-cancer classification task from our 2014 entry, and applied for the whole ICGC dataset, for samples with SSM (simple somatic mutation) data, this division results in a set of 3491 examples for complete remission and 1307 examples for progression until death.

In optimizing the parameters powers of ten in the range -10 to 10 are evaluated for the C-parameter of the SVM and values 10, 100 and 1000 are tested for the number of trees in the Extra Trees Classifier. In previous experiments we have seen performance increases when using up to 10,000 trees with ensemble methods but due to the large size of the cross-cancer datasets this is not feasible in the current experiments.

3.3 Features

Unlike our entry from 2014, in this work we use only one ICGC data type per classification experiment. This is primarily due to the number of cases where only some of the data types are available for a patient. For example, the SSM data is available for 7,908 patients whereas the EXP-A data is available for only 3,135 patients.

The primary data type used in our experiments is the SSM (simple somatic mutations). When generating features based on SSM the primary challenge is the sparsity of the data. Even the most common SSM in the ICGC data, MU62030 (a single base A>T substitution in the gene BRAF), occurs in only 405 donors across all the ICGC projects. Therefore, individual mutations have to be grouped if they are to be used as machine learning features. We first experimented with simply grouping all mutations within one gene, that is, we used simply the binary mutation status of a gene as a feature. While reaching decent classification performance, such features are not very interesting for the analysis of the mutational basis behind a certain classification. Therefore, in the final experiments we grouped mutations both by gene and by their functional impact on that gene (e.g. exon variant, intron variant, missense etc). This feature type mostly preserved the classification performance of gene-level features, while providing a more interesting feature set for the subsequent mutation analysis.

As a point of comparison for the SSM mutations we tested gene expression levels. Gene expression is commonly used a sort of "fingerprint" for the phenotype of a particular cancer. In a machine learning context gene expression levels are easier to work with than the SSM, as at least some value is present for each gene in each analysed sample, but the expression features are of course "one step removed" from the underlying genetic causes of the cancer. For the expression data we chose the sequencing based expression (EXP-S) as that is more commonly available for the ICGC samples than the array based expression (EXP-A).

3.4 Feature Analysis

The feature analysis is based on the embedded feature importance ranking provided by the Extra Trees Classifier [Breiman et al.]. In ensemble methods the relative rank (depth) of a feature contained in a decision tree can be used as a measure of the importance of that feature in performing the classification. In our 2014 CAMDA entry we have shown that compared with e.g. greedy forward selection and recursive feature elimination the embedded feature importance estimation results in relatively stable performance progression. To determine the relevance of the selected features we compare them against the genes in the COSMIC Cancer Gene Census.

4 Results and Discussion

4.1 Classification Performance

The classification results are shown in Table 1. The primary feature set of SSM results in a decent performance of slightly above 0.7 AUC. Both the support vector machine (SVM) as well as the extra trees classifier (ETC) provide similar performance. Unlike in our 2014 entry using the ETC does not result in higher performance, perhaps due to the larger class sizes. SSM-based classification with the SVM is generally faster and has slightly higher performance, but does not provide the embedded feature analysis. With both the SSM and EXP-S feature sets we observe a notable increase in performance compared with the five-fold cross-validation of the training set and the final classification of the hidden set. We speculate this may be due to the size of the datasets and the additional 20% of training data available when classifying the hidden set. A learning curve experiment should be done in the future to evaluate this assumption.

4.2 Feature Analysis

As seen in Figure 1 known cancer genes are more common among the features selected as most important. This alludes to some biological relevance behind the automatically learned classification and the automatic selection of features.

Table 2 shows the top 20 features from the feature importance ranking produced with the extra trees classifier for the SSM feature set. The most important feature turns out to be any mutation in an intergenic region. As such a feature has no gene name it becomes very common and is possibly slightly correlated with one of the two classes.

The more interesting features are those generated for mutations within known genes, as they define both a gene name and a functional consequence (depending on the mutated site). The top three genes, EGFR, KRAS and TERT are traditional, well known cancer genes. Mutations in the *epidermal growth factor receptor* (EGFR or ErbB-1) can lead to uncontrolled cell division and as such it is a central gene in a number of cancers [Lynch et al., 2004]. The *V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog* (KRAS) is a regulator of growth-related signaling and its mutation is essential for the growth of many tumours [Kranenburg, 2005]. Mutations in *Telomerase reverse transcriptase* allow telomerase to remain active in somatic cells and thus leads to immortal cancer cells [Zhang et al., 1999].

Table 1: Classification performance. Example counts are shown for the two classes followed by the AUC_T and AUC_H scores for the (t)raining and (h)idden sets.

features	classifier	remission	progression	AUC_T	AUC_H
SSM	Extra Trees	3491	1307	0.612 ± 0.056	0.704
SSM	Linear SVC	3491	1307	0.611 ± 0.030	0.722
EXP-S	Extra Trees	4592	1210	0.565 ± 0.033	0.819
EXP-S	Linear SVC	4592	1210	0.602 ± 0.044	0.758

Table 2: The most important features. Each feature is the id of the gene combined with the mutation consequence. The census column indicates whether the gene is among the known cancer genes in the COSMIC census.

#	gene id	gene name	consequence	census
1			intergenic region	
2	ENSG00000146648	EGFR	missense variant	•
3	ENSG00000133703	KRAS	missense variant	•
4	ENSG00000164362	TERT	upstream gene variant	•
5	ENSG00000121879	PIK3CA	missense variant	•
6	ENSG00000175826	CTDNEP1	stop gained	
7	ENSG00000023516	AKAP11	missense variant	
8	ENSG00000187172	BAGE2	intron variant	
9	ENSG00000141510	TP53	intron variant	•
10	ENSG00000169031	COL4A3	intron variant	
11	ENSG00000096968	JAK2	intron variant	•
12	ENSG00000182185	RAD51B	intron variant	
13	ENSG00000149531	FRG1B	stop gained	
14	ENSG00000141510	TP53	exon variant	•
15	ENSG00000115896	PLCL1	intron variant	
16	ENSG00000210154	MT-TD	downstream gene variant	
17	ENSG00000174473	GALNTL6	intron variant	
18	ENSG00000229981	LINC01435	intron variant	
19	ENSG00000130226	DPP6	intron variant	
20	ENSG00000140945	CDH13	intron variant	

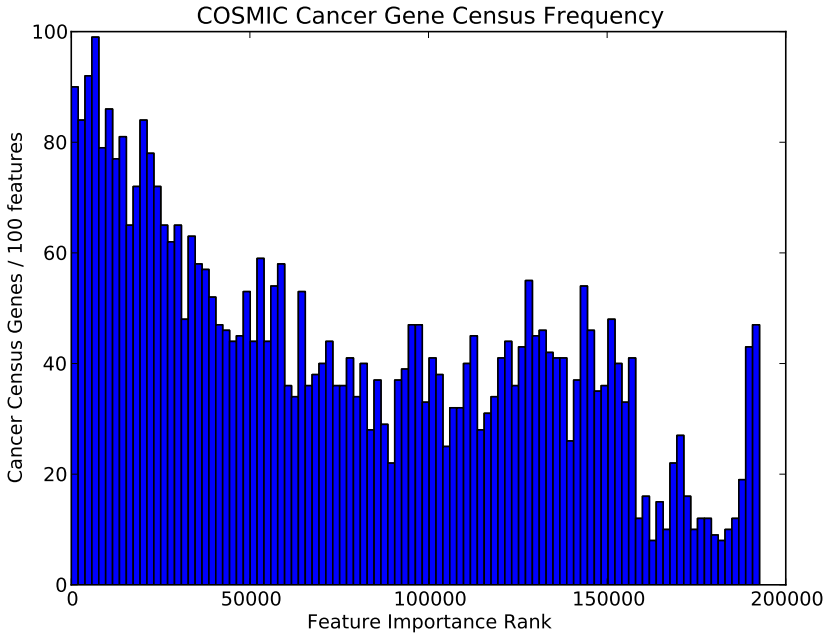

Figure 1: Known cancer genes among the selected features. Known cancer genes from the COSMIC census are more common among the features automatically selected for classifying cancer progression.

5 Conclusions

We have extended our classification-based cancer analysis approach to the entire ICGC cross-cancer dataset. With the increased size of the dataset, sparse feature groups such as the SSM become usable as classification features, and can achieve decent classification performance for predicting cancer progression. The SSM represents the most causally relevant feature set for understanding the nature of the ICGC cancers, as these individual mutations form the driving force of many tumours. While analysis of the feature selection results shows a correlation with known cancer genes, more work is needed to evaluate the role of the less known mutations.

The classification into complete remission or progression until death is a classification where all examples are positive for being cancers. However, common cancer genes present in the COSMIC census rank highly as features relevant for predicting the progression of cancer. We speculate that genes commonly mutated in cancer are also among the strongest drivers of cancerous growth, making them good indicators for the severity of progression, with mutations in several such genes being more likely to result in a fatal cancer.

As future work we hope to find ways to better utilize the mutation data on the level of individual mutations, to provide the kind of analysis required for the current biomedical research of separating the important driver mutations from the random passenger ones. As with our earlier project, we will publish all of our experimental code under an open source license[3].

References

J. Björne, A. Airola, T. Pahikkala, and T. Salakoski. Classification and feature selection across the ICGC Cancer Projects in the CAMDA 2014 Challenge. 2014.

L. Breiman, J. Friedman, and R. Olshen. Stone, cj (1984) classification and regression trees. *Wadsworth, Belmont, California*.

P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, 2004.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

O. Kranenburg. The kras oncogene: past, present, and future. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1756(2):81–82, 2005.

T. J. Lynch, D. W. Bell, R. Sordella, S. Gurubhagavatula, R. A. Okimoto, B. W. Brannigan, P. L. Harris, S. M. Haserlat, J. G. Supko, F. G. Haluska, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139, 2004.

J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, and A. Kasprzyk. International cancer genome consortium data portal – a one-stop shop for cancer genomics data. *Database*, 2011, 2011.

X. Zhang, V. Mar, W. Zhou, L. Harrington, and M. O. Robinson. Telomere shortening and apoptosis in telomerase-inhibited human tumor cells. *Genes & development*, 13(18):2388–2399, 1999.

[3] `https://github.com/jbjorne/CAMDA2015`