Write-up for BeatAML Dream Challenges

Team: CompOnc
Asimomitis Georgios[1], Bernard Elsa[1], Tazi Yanis[1]
[1]Computational Oncology, Memorial Sloan Kettering Cancer Center

We agree to make our submission public as part of the challenge archive.

SC1. Ex-vivo drug sensitivity prediction with kernel multi-task regression.

Abstract

We used kernel-based feature representation and multi-task learning to predict ex-vivo drug sensitivity from patient gene expression profiles by sharing information across drugs.

Introduction

Precision oncology seeks to select optimal therapeutic options for each patient based on the molecular characteristics of their disease. With scarce therapies available for AML patients and disparate responses to treatment, the identification of novel biomarkers of drug sensitivity is critical.

The BeatAML dream subchallenge 1 aimed to predict ex-vivo sensitivity of 122 drugs (cytotoxic or targeted agents) from genomic variants, gene expression or clinical variables of patients with AML. To leverage both the correlations among drugs and the flexibility of kernel methods, we implemented a kernel multi-task regression model.

Methods

<u>Feature selection</u>
Selecting the most informative features is a critical step for predicting the ex-vivo drug sensitivity. The predictive ability of the clinical and molecular features as well as their combination was assessed in a cross-validation framework across different predictive models (to include kernel-based or random forest models). Gene expression data only led to high and robust predictive performance. In particular, collecting the ~1000 genes with the highest average correlation with the AUC profiles across all 122 drugs, together with the top most correlated genes specifically per drug created a highly informative set of features. As kernel methods can operate in high-dimensional space, this set was also enriched with the top 1000 genes with highest variance across patients. The specifics of the feature sets are presented in Fig. 1B.

Kernel multi-task regression (KMR)

Kernel ridge regression is a state-of-the art method where a ridge regression is performed in a feature space derived by a kernel function [1].

Suppose we have n patients described by p descriptors (e.g. profile of gene expression), so that $x_i \in R^p$. Let y^k be an output vector for task k, i.e. the sensitivity of drug k.
The objective function of the regularized least square is $||y^k - f(x)||_2^2 + \lambda||w||_2^2 $ where the prediction model is $f(x) = \phi(x)^Tw + w_0$ with w the model parameters and w_0 a bias term, and where \lambda is a regularization parameter. A kernel representation of f(x) can be written as:

$$f(x) = \sum_{i=1}^{n} K(x_i, x)\alpha_i + w_0.$$

Our KMR method is a simple extension of kernel ridge regression where f is a bilinear function

Kernel computation

To utilize the strong patterns of correlations observed among drugs (Fig. 1A), the drug kernel Q was defined as the nearest positive semi-definite matrix of the correlation matrix [2].

For the choice of the patient kernel QK, i.e. similarity matrix between patients, we computed linear kernels from gene mutation and clinical data, and gaussian kernels from gene expression data. As also introduced in the previous section, we observed in cross-validation experiments that the former kernels had minimal and non-robust predictive power, so that we only selected gene expression for further modelling.

Implementation

Our final workflow is illustrated in Fig. 1B. We used the KMR implementation available at https://github.com/jpvert/kmr [3]. In practice, we trained a different KMR for each drug to allow for drug-specific regularization parameters.

We also trained a random forest individually for each drug (not multi-task) and compared its cross-validation predictive performances with KMR. For a few drugs (A674563, Tandutinib, Volasertib) where random forest was significantly superior to KMR we switched the predictive model (Fig. 1B).

Discussion

Our KMR method utilizes the correlations among drugs to share information between predictive tasks in the learning process. It integrates kernel-based feature representation (gaussian kernel of gene expression) and multi-task learning (empirical drug correlation kernel). Reassuringly, our best cross-validation performance was achieved for Venetoclax (Fig. 1A), which has been recently FDA-approved for newly diagnosed AML patients older than 75 years old. Future work might consider

the incorporation of drug features, such as structural descriptors or interactions with target proteins, in the multi-task framework.

SC2: Outcome prediction from knowledge transfer of large scale AML genomic study.

Abstract

To predict overall survival of AML patients, we augmented the BeatAML training data with a large-scale AML publicly available genomic study and implemented an ensemble approach that combined risk predictions learned from different datasets.

Introduction

Therapeutic oncologic decisions are informed by the estimate of risk for each patient. Accurate risk estimation is critical in AML: higher risk patients are considered for allogeneic hematopoietic cell transplant, which is associated with >20% mortality rate and chronic grafts-versus-host disease.

The BeatAML dream subchallenge 2 aimed to predict the relative risk of AML patients based on ex-vivo drug sensitivity data, genomic variants, gene expression and clinical data. Importantly, large scale translational genomics studies have identified the genes recurrently mutated in AML and showed their prognostic relevance [4] [5] [6].

We implemented a hybrid approach that borrowed concepts from transfer and ensemble learning. Our 2-step method learned two predictive models: i) one model solely based on the BeatAML clinical and genomic data and ii) another model built from an augmented dataset integrating publically available AML clinical and genomic data [5] [6]. Our approach ultimately combined the two independent predictions as an optimized weighted-average of the rank of the estimated patient risks.

Methods

Feature definition

We selected known clinically relevant variables, such as age, bone marrow blasts, white blood cell count as well as therapy-related or secondary neoplasms.

The most frequently mutated genes in AML have been extensively described [4] [5]. The BeatAML genomic data replicated these findings with high frequency of *NPM1, FLT3, NRAS, TET2, DNMT3A, SRSF2, IDH1/2* and *TP53* mutations among others. We excluded some genes from the set of recurrently mutated genes in BeatAML data (*ZNF711*, *DDX60L* and *TPRX1*), as examining the variants of those genes pointed towards non-pathogenicity. In total, we selected 25 genomic descriptors (Fig 2A).

In addition to clinical and genomic descriptors, we tested the prognostic relevance of ex-vivo sensitivity and gene expression data. As expected for experimental drugs tested on patient derived cell lines, the predictive ability of ex-vivo sensitivity data for overall survival was close to zero. Surprisingly, although gene expression data had good internal cross-validation results, generalization ability on the leaderboard data for this task was weak, so that we only further considered clinical and genomic descriptors.

Augmented training data

We used an external dataset of 1,540 AML patients containing molecular and clinical information as well as outcomes, and extracted common features with the challenge dataset. We further evaluated feature importance from cross-validation and selected 16 common descriptors between the beatAML and external dataset (2 clinical and 14 molecular, Fig 2A).

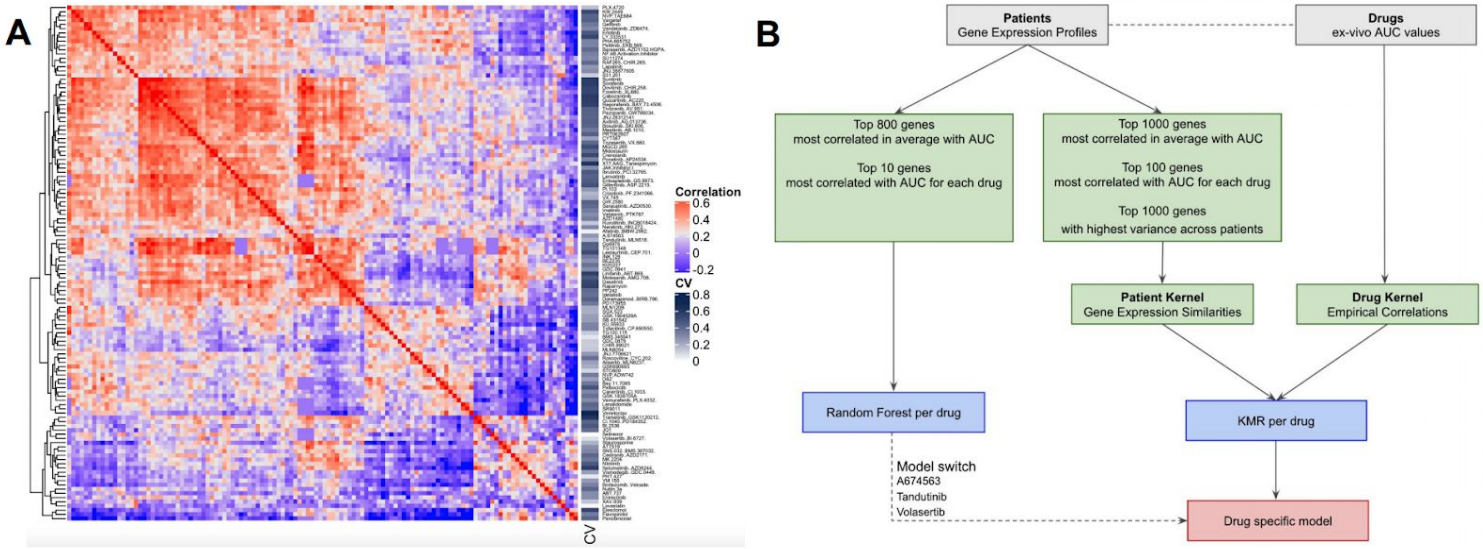
Ensembl model with weighted average of risk ranks

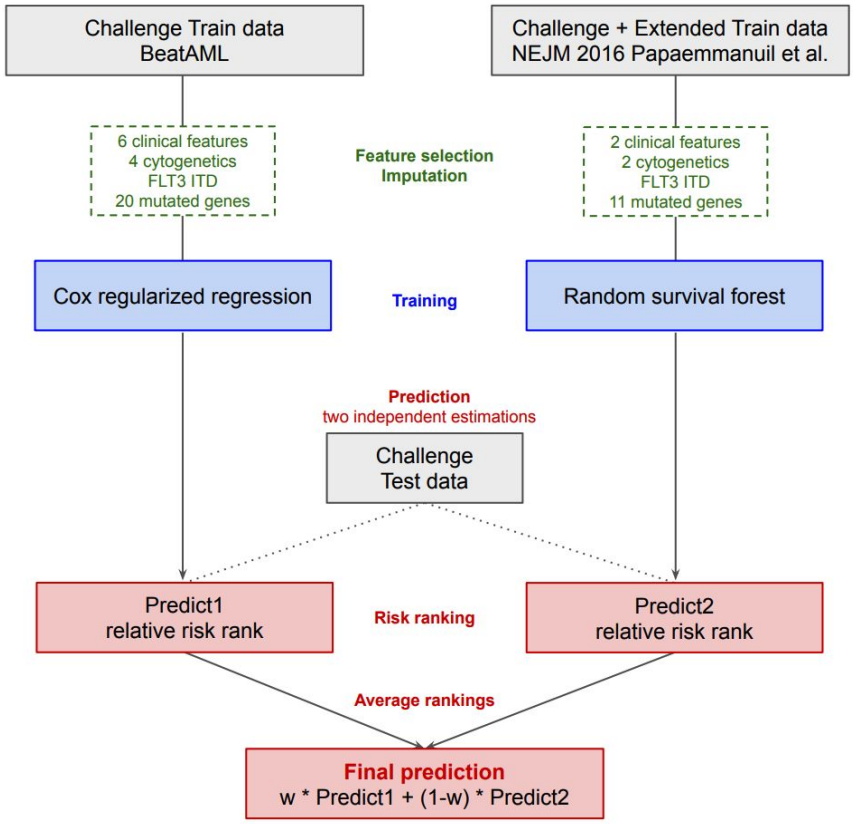
As we initially observed that cross-validation prognostic performances increased with the incorporation of the augmented dataset (Fig 2B?), we implemented an ensemble approach that first learned two independent models on either the beatAML data or on the augmented dataset. The choice of each predictive model (Cox regularized regression and random survival forest) was determined by cross-validation (Fig. 2A).

The predicted risks of the two models were then ranked, and our final prediction consisted of a weighted-average of the two vectors of relative rank risks (Fig. 2A). The weight was also determined with cross-validation.

Discussion

We augmented the beatAML training data with a publicly available AML study and we combined the predictions of two independent models trained on different training data. This work highlights that molecular and clinical information can be shared across datasets and that prognosis modeling strongly benefits from publicly available datasets.





SC2 Workflow

Prognosis added value from external data across models

References

(suggested limit 10 references)

1. Schölkopf B, Smola AJ, Managing Director of the Max Planck Institute for Biological Cybernetics in Tubingen Germany Profe Bernhard Scholkopf, Bach F. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press; 2002.

2. Knol DL, ten Berge JMF. Least-squares approximation of an improper correlation matrix by a proper one. Psychometrika. 1989. pp. 53–61. doi:10.1007/bf02294448

3. Bernard E, Jiao Y, Scornet E, Stoven V, Walter T, Vert J-P. Kernel Multitask Regression for Toxicogenetics. Mol Inform. 2017;36. doi:10.1002/minf.201700053

4. Cancer Genome Atlas Research Network, Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013;368: 2059–2074.

5. Papaemmanuil E, Döhner H, Campbell PJ. Genomic Classification in Acute Myeloid Leukemia. The New England journal of medicine. 2016. pp. 900–901.

6. Gerstung M, Papaemmanuil E, Martincorena I, Bullinger L, Gaidzik VI, Paschka P, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat Genet. 2017;49: 332–340.

- Elli Papaemmanuil, Moritz Gerstung, Peter J. Campbell et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia The NEW ENGLAND JOURNAL of MEDICINE. 2016
- Elsa Bernard, JP Vert et. al, Kernel Multitask Regression for Toxicogenetics , 2017

- Ping Wang , Yan Li , Chandan K. Reddy. Machine Learning for Survival Analysis: A Survey. 2017
- Hemant Ishwaran Random Survival Forests. The Annals of Applied Statistics. 2008
- Park M, Hastie T. l1-Regularization Path Algorithm for Generalized Linear Models. Journal of the Royal Statistical Society. 2007

Authors Statement

All authors contributed equally