TOPSPIN: a novel algorithm to predict treatment specific survival in cancer

Joske Ubels[1,2,3], Erik H. van Beers[3], Pieter Sonneveld[2], Martin H. van Vliet[3] and Jeroen de Ridder[1]

[1]Department of Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands, [2]Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands, [3]SkylineDx, Rotterdam, The Netherlands

Introduction

It is increasingly recognized that the successful treatment of cancer is hampered by genetic heterogeneity of the disease and a more personalized approach is needed. Differences in the genetic makeup between tumors can result in a different response to treatment (Burrell et al., 2013). As a result, despite the existence of a wide range of efficient cancer treatments available (Block et al., 2015), many therapies only benefit a minority of the patients that receive them, while they are associated with very serious side effects. Therefore, there is a great clinical need for tools to predict - at the moment of diagnosis - which patient will benefit most from which treatment. This requires the discovery of markers, like gene expression signatures, that are informative about treatment response.

The first gene expression signature that was shown to successfully predict prognosis in cancer was a 70-gene breast cancer signature (Van 't Veer et al., 2002). In recent years, many more gene expression signatures that can distinguish molecular subtypes of cancer or predict a favourable prognosis have been published, for a wide variety of cancers. However, by definition, prognostic signatures predict survival irrespective of the treatment given. To aid in treatment decision we need to discover a predictive signature, i.e. a marker that can predict survival depending on which treatment is given.

Building a classifier for predictive purposes poses a unique challenge, which is not encountered when inferring prognostic classifiers. Most methods for defining a prognostic classifier rely on a supervised learning approach. In these methods a label is defined for each patient based on their survival or some other outcome measure, like the risk of experiencing a relapse. The training procedure then focuses on predicting these labels as accurately as possible to ultimately produce a classifier that can predict outcome for a new patient. However, when building a predictive classifier we aim to predict whether a patient will have a better prognosis when given a certain treatment of interest as compared to a different treatment. This means that labels defined solely on survival data will be inadequate, since it is impossible to know whether a patient would have had a different outcome when given an alternative treatment. A patient with a favourable outcome when given the treatment of interest, may have responded as well to any other treatment. Conversely, a patient may have a shorter than average survival even when treated with the optimal treatment; any other treatment would have resulted in an even worse prognosis. The absence of predefined labels make existing methods for building gene expression signatures unsuitable for this problem and thus a novel approach is needed.

To address this challenge we introduce a new algorithm, TOPSPIN (Treatment Outcome Prediction using Similarity between PatIeNts), that derives a classifier able to distinguish a subset of patients with improved treatment outcome from the treatment of interest, but not the comparator treatment. Uniquely, TOPSPIN integrates the process of defining labels and building a classifier, eliminating the need to predefine labels based on survival alone. The fundamental idea of our approach is that we can estimate a patient's treatment benefit by comparing its survival to a set of genetically similar patients that received the comparator treatment. Patients with a large survival difference can then act as prototype patients: new patients with a similar gene expression profile should also benefit from receiving the treatment of interest. These prototype patients are simultaneously used to define the classifier and the labels.

In this work we focus on Multiple myeloma (MM), which is a clonal B-cell malignancy that is characterized by abnormal proliferation of plasma cells in both the bone marrow and the extramedullary sites. Median survival is 5 years (Howlader, 2016). In the last two decades many novel therapies have been introduced for MM, resulting in an improved survival. However, response rates remain low and there is no clear indication what determines treatment response. This is complicated by the fact that MM is very heterogeneous, both between and within patients (Lohr et al, 2014). Especially in MM, predictive signatures could be of great benefit.

Methods

Data

We pooled gene expression and survival data from three phase III trials: Total Therapy 2 (TT2, GSE2658), Total Therapy 3 (TT3, GSE2658) and HOVON-65/GMMG-HD4 (H65, GSE19784). In our analyses of the pooled data two treatment arms were considered: a bortezomib arm, which comprises the PAD arm from H65 and TT3, and a non-bortezomib arm, which comprises the VAD arm from H65 and TT2. Combined, these datasets include 910 patients, for which 407 received bortezomib and 503 did not. We split the dataset in a training set (n = 606) and a test set (n = 304). This test set is not used at any point in the training procedure and acts as an independent validation set to assess the performance of the final classifier.

Progression free survival (PFS) was used as outcome measure.

Algorithm

TOPSPIN aims to predict if a patient benefits or does not benefit from a certain treatment of interest based on the gene expression profile of the patient. In order to train this classifier, we split the training set into three equal folds (A, B & C). We first define a ranked list of prototype patients on fold A (Step 1) that exhibit a better than expected prognosis compared to a set of genetically similar patients that received the opposite treatment. In Step 2, a decision boundary around a selection of prototype patients is determined on fold B. Patients who lie within this decision boundary are expected to show a favourable outcome when receiving bortezomib and and make up class F. All other patients are considered class N and are not expected to benefit from receiving bortezomib. Because it is a priori unknown based on which genes patient similarity should be defined, Step 1 and 2 are performed for a large number of functionally coherent gene sets obtained from the Gene Ontology annotation, yielding one classifier per gene set. Step 1 and 2 are repeated *k* times for all *n* gene sets, which ultimately results in *n* * *k* classifiers. In an approach based on the boosting principle, the individual classifiers are combined to construct a more robust final classifier. These classifiers are applied separately to the samples in fold C, which act as out-of-bag samples. Since across the repeats all samples are included in fold C this will give an independent classification per gene set for all patients included in the training dataset. The performance of a classifier is defined by the Hazard Ratio (HR) found between the two treatment arms within class F. Since not all the trained classifiers will be equally successful in identifying the subset of patients that benefits from the treatment of interest a threshold S is set in Step 3, which determines which classifiers will participate in the final classifier: a classifier is included only if its performance is below a certain HR. We base this threshold on a 50/50 mixture of the performances obtained on fold B and fold C, the OOB samples. This defines a binary vector **x** for each patient, where x_s has a value of 1 if the patient belongs to class F according to the s^{th} classifier and 0 otherwise. A classification score is defined for a patient *i* based on **x**:

$$Classification\ Score_i = \frac{\sum_1^s x_s}{s}$$

with *s* being the number of classifiers contributing to the final classifier. On this score a threshold T is set, which determines whether a patient is to benefit from the treatment of interest. These steps are visualized in Figure 1 and are described in more detail below.

<u>Step 1 - Prototype ranking on fold A</u>

For each patient receiving the treatment of interest, the treatment benefit is defined as

$$\Delta PFS_i \ = \frac{1}{n}\sum_{j \in O}(PFS_i - PFS_j)\,,$$

where O is the set of the *n* most similar patients (based on euclidean distance) that did not receive the treatment of interest. We use *n* = 10. ΔPFS is only calculated for neighbor pairs where it is clear which patient experienced an event first; if both are censored, ΔPFS is not computed. To correct for the fact that a patient with a long

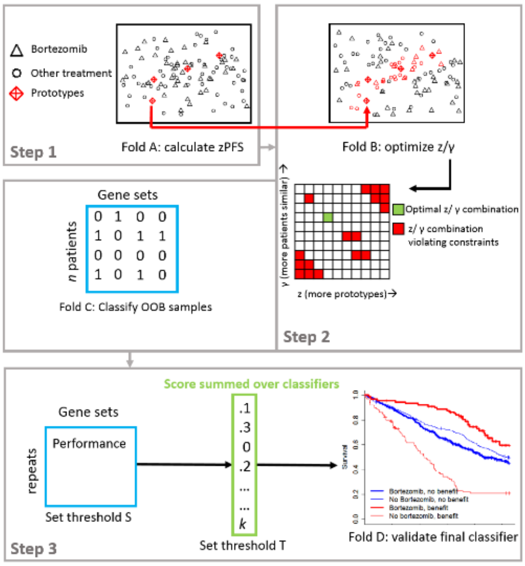

Figure 1. Overview of the TOPSPIN algorithm

survival time will, on average, have a large ΔPFS irrespective of its relative treatment benefit compared to genetically similar patients, we define the z-normalized zPFS score as:

$$zPFS_i = \frac{\Delta PFS_i - \mu(RPFS_i)}{\sigma(RPFS_i)},$$

where RPFS is a distribution of 1000 random ΔPFS scores, obtained by calculating ΔPFS for randomly chosen sets O, i.e. determining treatment benefit with respect to random patients instead of genetically similar patients. Based on the zPFS score all patients in fold A that were given the treatment of interest can be ranked.

Step 2 - Classifier definition on fold B

Classifier Q is defined by a subset of z top-ranked prototypes along with a decision boundary defined in terms of the euclidean distance γ around a prototype. A patient is classified as class F when it lies within γ of any of the top z prototypes. The optimal values for z and γ are those resulting in the lowest Hazard Ratio (HR) in class F (the patient group in which the treatment of interest should have a better survival). We additionally constrain z and γ, such that class F comprises at least 20% of the dataset. The number of prototypes was restricted to 3 to prevent defining an extremely complicated classifier. The search grid for parameter γ was made dependent on the local density of the neighbors, and consisted of the sorted list of euclidean distances between the prototype and its neighbors. The optimal z and γ combination is chosen so that the HR in class F is minimal, with a preference for a HR associated with a p-value < 0.05.

Step 3 – Set thresholds S and T

A threshold S which determines which classifiers are included in the final classifier is optimized. Any classifier that resulted in a HR higher than S is excluded, with the options ranging from 1 to 0.3 in steps of 0.025. To utilize the information gained in Fold C, but prevent overtraining, the performance used is alternately the HR found on fold B and the HR found on fold C. For each possible threshold S, a threshold T is also optimized. This threshold T is set on the Classification Score to define class F in the final classifier. The combination of S and T that leads to the HR associated with the lowest p-value in class F, given that class F comprises at least 20% of the dataset, is chosen.

Results

The optimal threshold S found was 0.45, with a threshold T of 0.3. With this threshold S in total 14 150 classifiers were included in the final classifier. Applying the final classifier resulted in a HR of 0.43 (p = $1*10^{-4}$) in the training set within class F, based on the classification of the OOB samples. More importantly, when applied to the independent test set, a HR of 0.5 (p = 0.04) between the two treatment arms was found, demonstrating TOPSPINs ability to identify the subset of patients benefitting from bortezomib. The Kaplan Meiers for the training and test classification are shown in Figure 2. It is important to note that in class N HRs of 0.9 and 0.97 were found in the training and test set, respectively. These patients did not experience any benefit from receiving bortezomib and could thus possibly have been spared the treatment and side effects.

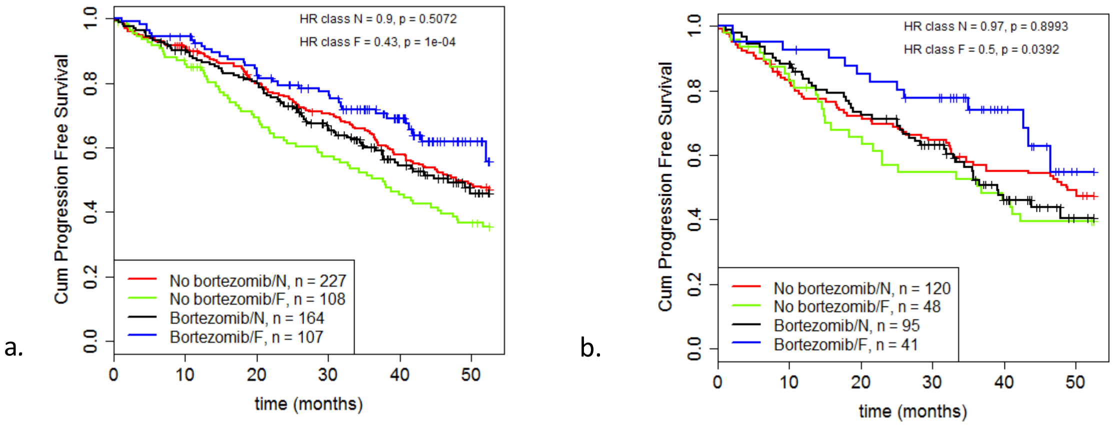

Figure 2. a. Kaplan Meier of the optimal classification of the training set. b. Kaplan Meier of the optimal classification of the test set.

Conclusion

Here we have demonstrated TOPSPIN ability to identify a subset of MM patients that benefit from the proteasome inhibitor bortezomib. TOPSPIN is however not specific for MM and can be used on any dataset with two randomized treatment arms and a continuous outcome measure. Considering the often low response rates combined with the serious side effects of current cancer therapies, TOPSPIN therefore offers an important step towards realistic personalization of cancer medicine.

References

Block, K. I., et al. (2015). Designing a broad-spectrum integrative approach for cancer prevention and treatment. Seminars in Cancer Biology, 35, S276–S304. doi:10.1016/j.semcancer.2015.09.007

Burrell, R. A., et al. (2013). The causes and consequences of genetic. Nature, 501, 338–345. doi:10.1038/nature12625

Howlader N, et al. (2016). SEER Cancer Statistics Review, 1975-2013. In National Cancer Institute. Bethesda, MD. Retrieved from http://seer.cancer.gov/csr/1975_2013/

Lohr, J. G., et al. (2014). Widespread genetic heterogeneity in multiple myeloma: Implications for targeted therapy. Cancer Cell, 25, 91–101. doi:10.1016/j.ccr.2013.12.015

Van ’t Veer, L. J., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(6871), 530–536. Retrieved from http://dx.doi.org/10.1038/415530a