FloReMi: Survival time prediction based on flow cytometry data

Sofie Van Gassen SOFIE.VANGASSEN@UGENT.BE
Data mining and modelling for biomedicine, VIB Inflammation Research Center, Ghent, Belgium

INTEC/IBCN - iMinds, Ghent University, Ghent, Belgium

Celine Vens CELINE.VENS@KULEUVEN-KULAK.BE
Department of Public Health and Primary Care, KU Leuven Kulak, Kortrijk, Belgium

Tom Dhaene TOM.DHAENE@UGENT.BE
INTEC/IBCN - iMinds, Ghent University, Ghent, Belgium

Bart N. Lambrecht BART.LAMBRECHT@UGENT.BE
Immunoregulation and Mucosal Immunology, VIB Inflammation Research Center, Ghent, Belgium

Department of Internal Medicine, Ghent University, Ghent, Belgium

Yvan Saeys YVAN.SAEYS@UGENT.BE
Data mining and modelling for biomedicine, VIB Inflammation Research Center, Ghent, Belgium

Department of Internal Medicine, Ghent University, Ghent, Belgium

Keywords: single cell analysis, flow cytometry, survival time prediction

Abstract

Flow cytometry data offers a detailed view on the immune system, but can be hard to mine for relevant biological conclusions. We applied an automated pipeline including preprocessing, feature extraction, feature selection and survival analysis to predict the progression time to AIDS for HIV patients based on flow cytometry data from blood samples. Our method selected a small subset of potential biomarkers relevant for AIDS progression and ranked first in the FlowCAP IV challenge.

1. Flow Cytometry

Flow cytometry is a high-thoughput technique for single cell analyses. To study blood or tissue samples, researchers can measure cell properties such as cell size, granularity and the presence of specific proteins on millions of individual cells. Current state-of-the-art flow cytometers measure up to 30 markers simultaneously, while mass cytometers go up to 50 and more.

Most researchers still explore these datasets in a manual way, by iteratively selecting cells in 2D scatter plots. This leads to results biased towards the expected cell types and many cells are discarded even though they might contain relevant information. Machine learning techniques can provide more objective solutions, as they are better suited to handle high-dimensional data.

2. The FlowCAP IV challenge

The FlowCAP initiative stimulates novel research in computational flow cytometry and provides multiple challenges to evaluate the performance of machine learning techniques on cytometry data. High-quality benchmark datasets have been made available for clustering and classification (Aghaeepour et al., 2013), and their most recent FlowCAP IV challenge (Aghaeepour et al., 2015) focused on a survival modeling challenge. Participants had to design a model to predict the time of progression from HIV to AIDS, based on flow cytometry data from a stimulated and unstimu-

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

lated blood sample. The main challenge in this dataset was the high censor level. This meant that for about 80 percent of the patients in the training dataset, it was not known when the progression to AIDS occured, but only a timepoint when it did not happen yet was given. The FloReMi approach using random survival forests, the method described below, ranked first of the nine methods submitted to the challenge.

3. FloReMi

FloReMi (Flow density survival Regression using Minimal feature redundancy) is a pipeline consisting of four parts: preprocessing, feature extraction, feature selection and survival time prediction. This algorithm is described in Cytometry Part A (Van Gassen et al., 2015).

Preprocessing of flow cytometry data is crucial to obtain good results. We applied quality control to remove timeframes with measurement problems, removed margin events which fall out of the detection range of the machine, removed doublets (multiple cells sticking together instead of single cell measurements), applied compensation to fix spillover from the multiple detection channels, applied logicle transformation (a well-known transformation function for flow cytometry data, which stays linear around zero and changes to a log scale for larger values) and finally selected live T cells based on the CD3 and Vivid/CD14 channels to zoom in on the cell types of interest.

Once the data is cleaned up, the feature extraction process can start. Instead of a simple feature vector for each patient, flow cytometry offers feature vectors for hundreds of thousands of cells for every patient. We used the flowDensity algorithm (Malek et al., 2015) to detect the splits between high and low feature values, and used the flowType algorithm (Aghaeepour et al., 2012) to group the cells into every possible combination of high and low feature values. For each of these groups, we calculated the percentage of cells and their mean feature values. As we had stimulated and unstimulated samples available, we applied this method on both and also computed the difference between the two. This resulted in a total of 2.480.058 features per patient.

Prediction algorithms cannot handle this amount of feature as-is, so the next step is a feature selection algorithm. We computed the p-value based on the Cox proportional-hazard model for every single feature, used this information to sort the features based on their correlation with the progression time, and selected features which did not correlate strongly with each other. This lead to a selection of 13 features.

Finally, we used the survival random forest algorithm (Ishwaran et al., 2008) on these 13 features to predict the progression time to AIDS. We also used the Cox proportional-hazards model with the same features and the additive hazards-model (with the 100 first sorted features), which gave similar results on the training data. However, their results were no better than random on the test data, indicating overfitting. As the survival forest is able to model non-linear relationships, we assume this gives it the final edge to make a good prediction.

Another challenge remains: the biological interpretation of the selected features, as the generated features do not correspond one-on-one with known cell types. It is likely that our best feature corresponds with effector CD8+ cells, but future research is necessary to gain more insight in these results.

Acknowledgments

SVG: Ph.D. Grant of the Agency for Innovation by Science and Technology in Flanders (IWT)
CV: Fund for Scientific Research Flanders (FWO)
Ghent University Multidisciplinary Research Partnership Bioinformatics: from Nucleotides to Networks.

References

Aghaeepour, N., et al. (2012). Early immunologic correlates of hiv protection can be identified from computational analysis of complex multivariate t-cell flow cytometry assays. *Bioinformatics*, *28*, 1009–1016.

Aghaeepour, N., et al. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, *10*, 228–238.

Aghaeepour, N., et al. (2015). A benchmark for evaluation of algorithms for identification of cellular correlates of clinical outcomes. *Cytometry Part A*, *89*, 16–21.

Ishwaran, H., et al. (2008). Random survival forests. *The annals of applied statistics*, 841–860.

Malek, M., et al. (2015). flowdensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics*, *31*, 606–607.

Van Gassen, S., et al. (2015). Floremi: Flow density survival regression using minimal feature redundancy. *Cytometry Part A*, *89*, 22–29.