An Ensemble of Classifiers for Time Course Classification of Response to Treatment in Psoriatic Patients.

Joel Correa da Rosa* Sandra Garcet †Jaehwan Kim‡ Suyan Tian§
James G. Krueger¶ Mayte Suárez-Farinas‖

Abstract

Psoriasis is caused by the complex interaction of genetic, environmental and immunological factors. However, psoriasis is also triggered by environmental factors such as certain medications, stress and recurring infection. In the last 10 years there has been an explosion of drug testing in Psoriasis with more than 35 new drugs in different phases of drug development. In this study the main goal is to build a genomic classifier, based on gene expressions measured at early time points (baseline, 1st, 2^{nd} and 4^{th} weeks) , that can predict accurately response-to-treatment in psoriatic patients at week 12 or later. To achieve this goal, we developed two classifiers: a universal that can be applied to any treatment and a treatment-specific. The first one predicts response-to-treatment independently of the drug and the second may reveal specificities related to the agent being used. Additionally, we are also interested in finding the earliest time point that allows us to predict response-to-treatment with desired accuracy. For the development of the classifiers we have transformed longitudinal skin expression profiles from 186 patients treated with 7 different drugs, and also placebo, into time course scores using LDA-projections. Gene Set Variation Analysis (GSVA) was applied for evaluating activity for some known psoriasis pathways as described in [4]. The final classification is obtained from an ensemble of 4 dissimilar classification methods: PLS (Partial Least Squares) , GLMnet (Generalized Linear Model), TGDR (Threshold Gradient Descent Regularization) and PAM (Predicition Analysis of MicroArrays). Bootstrapped samples were selected for simultaneously estimating tuning parameters and bagging genes in order to select stable gene signatures according to the observed bagging frequency (see [8]). The results pointed out that accuracy is increasingly gained as more time points are incorporated into the time-course score and very high accuracy ($> 95\%$) is obtained for the universal classifier at the second (2^{nd}) week of treatment.

1. Introduction

The development of drugs targeting key modulators of immune function has brought about a revolution in the treatment of many diseases. However, such drugs pose unique and novel challenges for drug development: as they target specific human immune components it is often harder to extrapolate from in vitro or animal models what their clinical efficacy may be, and the clinical response may require months to manifest. Thus longer trials with human patients are necessary, adding considerable cost, complexity and development time to the current cost of between 1.5 to more recently 2.6 billion dollars and 8 years in development for each successful drug brought to market (see [5] and [6]).

Psoriasis, a skin disease caused by the complex interaction of genetic environmental and immunological factors has become the success story of translational science. In the last decade, availability of high-throughput technologies for comprehensive molecular

*The Rockefeller University, Laboratory of Investigative Dermatology, New York
†The Rockefeller University, Laboratory of Investigative Dermatology, New York
‡The Rockefeller University, Laboratory of Investigative Dermatology, New York
§Division of Clinical Epidemiology, First Hospital of the Jilin University, Changchun, Jilin, China
¶The Rockefeller University, Laboratory of Investigative Dermatology, New York
‖The Rockefeller University, Laboratory of Investigative Dermatology, New York, Icahn Mount Sinai School of Medicine

characterization of the disease have greatly increased our understanding of psoriasis, and multiple biological therapies have proved to be highly effective.

Despite the increase in FDA-approved treatments for psoriasis, 20-30% of patients still fail to respond to biologics ([10]) and their effectiveness in real-world practice is lower than that reported in trials ([3]). Even specialized dermatologists have difficulties in selecting treatment for individual patients as to maximize the probability of clinical response ([1],[11]). Physicians recommend treatment options based on results of randomized clinical trials (RCT) and evidence syntheses, such as meta-analysis ([9]).

The ideal situation would be a test that could predict which treatment would be most likely to be beneficial for an individual patient, that is "personalized medicine". In fact, there is currently a tremendous interest in personalized medicine and on finding biomarkers to find drug and/or dose to predict response-to-treatment on individual patients.

Some attempts have been made on developing classifiers to treatment in psoriasis. In [7] the authors used baseline expression profiles from peripheral blood mononuclear cells (PBMCs) of 16 patients to develop a 23-genes signature that predicts week 12 response to Alefacept with 86% accuracy. More recently [2] showed that HLA-Cw6+ patients respond faster and hold response longer than HLA-Cw6- patients. However this study just looks at the association between response and genotype and does not develop a rule to predict response, a more challenging problem.

We built a setup for estimating the efficacy profile of new psoriasis drugs from a short-term proof-of-concept study. This prediction tool, uses skin biopsies obtained in a short-term (4 weeks) clinical trial to evaluate the efficacy profile of a psoriasis drug after a full treatment course ($>$12 weeks). Using longitudinal gene expression profiles obtained at baseline and intermediate time-points during this short-term trial, the tool can predict the percentage of patients in the cohort that will responde to treatment at the unobserved end-points of 12 weeks.

2. Methods

To develop the classifier, we have used longitudinal skin expression profiles from 186 patients treated with Etanercept, Ustekinumab, Adalimumab, Methotrexate and Placebo. These data have been either previously published or shared by collaborators; they include phenotypic information and microarray gene expression data (obtained using Affymetrix and Illumina platforms) of lesional and non-lesional skin of psoriatic patients taken at baseline, and after 1,2 and 4 weeks of treatment. The response to treatment was measured by the PASI75 endpoint where patients are considered responders if there was an improvement of at least 75% on the PASI score from baseline and non-responders otherwise.

In the development of our classifiers we took into account two major desirable characteristics of any predictive rule that aims to be extended successfully into diagnostic practice: robustness and platform independence. Robustness is achieved by careful consideration of potential sources of bias and by using adequate methodology.

The theory of high dimensional classification is now well developed for unstructured or hierarchically structured data. Our mathematical challenge is that we have far more genes than patients, and more patients than time-points. Our first step therefore is to select a space of classification algorithms which are methodologically appropriate for this situation. The regularization algorithms are able to deal with this problem by performing feature selection and model estimation simultaneously, and their development and application has exploded in the last two decades. Our second step was to build robustness as well as internal controls into our methods by choosing four very dissimilar regularization algorithms, chosen as to maximize the Jackcard dissimilarity metric. By combining the output of all four methods

we achieve higher robustness (by a "wisdom of crowds" principle) while by comparing the four outputs we get a means to check the consistency of the predictions. The chosen algorithms are : Partial Least Squares (PLS) , Prediction of MicroArrays (PAM), Generalized Linear Model with elastic net (GLMnet) and TGDR (Threshold Gradient Descent Regularization).

But additionally in our case the sucession of time-points is meaningful, requiring a time-series analysis approach. However, time series analysis methods are well developed for the case where there are far more time-points than measured variables and subjects what is not realistic with high-throughput data. We therefore modified all four algorithms to account for the within-patient correlation by transforming the multivariate time-course profile into an univariate time-course score. This is achieved by using a linear combination of time-course gene expressions that maximizes the variability between responders and non-responders relatively to the within-variability (LDA-projection) . These projections can be used with most of the regularization algorithms that are available today in public tools like weka , R or python, and in particular with our chosen algorithms. To emphasize that algorithms are applied to time-course scores we use the acronyms T-PLS, T-PAM, T-GLMnet and T-TGDR.

For each algorithm, model estimation was carried out using all available training data and cross-validation was used to select the optimal tuning parameters. Bootstrap simulation (N=500) were run to study the robustness of the classifier and the sensitivity of its error rate. Additionally, we couple each classifier with bagging technique to evaluate the robustness of the selected predictors (genes and pathways), which is known to improve classifier performance. As such, bootstrapped samples were used for simultaneously estimating tuning parameters and bagging genes in order to select a more stable gene signature.

Since gene-expression profiles are dependent on the technology used, we also developed a platform-independent classifier by using pathway-based profiles, known to have better reproducibility across platforms. Besides our gene-based time-course score, we will use Gene Set Variation Analysis (GSVA) technique to create pathway-sepecific time-course profiles for each patient and proceed with the same strategy to generate the pathway based classifiers. The general strategy for building the classifier is depicted on Figure 1.

Besides feasibility, we would like to determine what is the shortest time course trial that can be used to achieve a good accuracy in response prediction for week 12 or later. For this goal, we applied the methodology for different time spans: 0 (baseline), 0-1 weeks, 0-1-2 weeks and 0-1-2-4 weeks.

To boost the performance of the individual classifiers, they were ensembled by averaging the probabilities of response-to-treatment computed by each of them. This follows the ' wisdom-of-crowds' principle that assures that diverse and accurate classifiers will provide better results when combined.

3. Results

The performance of the universal and treatment-specific classifiers as measured by percentual accuracy in prediction are shown in Figure 2. Accurarcy is increasingly gained as more time points are incorporated into the computation of the time-course score while variation in accuracy decreases indicating that the classifiers are more stable.

Except for T-PLS algorithm, bagging improves the performance of the gene-based universal classifiers while reducing the number of genes in the signature by half when data up to weeks 2 and 4 are used. The improvement is marked for T-PAM. The same trend is observed for the treatment-specific classifiers with up to 8% improvement.

These pictures draw a very positive scenario where we can conduct treatment specific

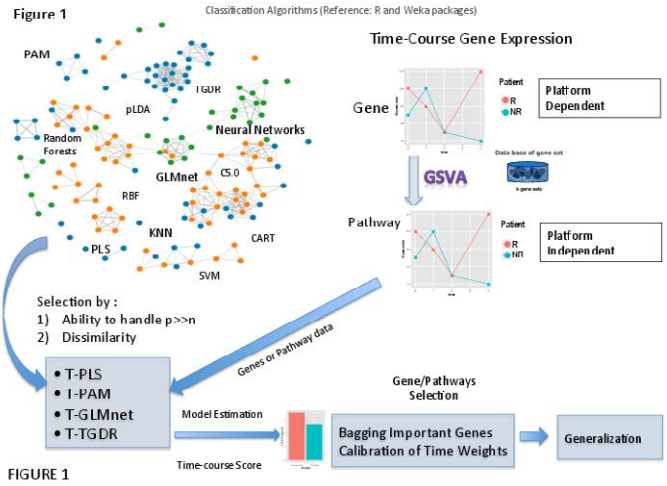

Figure 1: Strategy for building a Platform-Independent classifier, Time-course gene expressions or pathway activity scores are used as inputs to 4 different classification algorithms: T-PLS, T-PAM, T-GLMnet and T-TGDR, selected from an universe. During training stage, parameters and time-course scores are simultaneously estimated. A boostrapping experiment is carried out to bag important genes, estimate tuning parameters and weights for the time course score

classifiers to attain an accuracy of more than 95% after 1 week of treatment with the ensemble classifier and an impressive ($> 97\%$ accuracy with T-PAM). Treating a single patient for two weeks and evaluate the response can be done with almost 100% accuracy.

When applying the classifier to unseen data (generalization) we could find results at the gene-level with very high sensitivity at low specificty with data up to weeks to and 4 as can be seen in Figure 3 A. At the pathway level, the results are not better than the ones obtained by using the PASI as the only predictor.

4. Conclusions

In this work, we developed a strategy to combine classifiers into a ensemble that was used to predict response-to-treatment in psoriatic patients. The strategy was applied independently of the drug (universal) and for each specific drug (drug-specific). The results pointed out that is possible to obtain accuracy higher than 95% based on gene expressions from baseline, first and second weeks. Bootstrap resampling combined with bagged genes yielded more stable and accurate predictions. Results out-of-sample confirmed the potential of this technique to have good performance in unseen data.

Figure 2

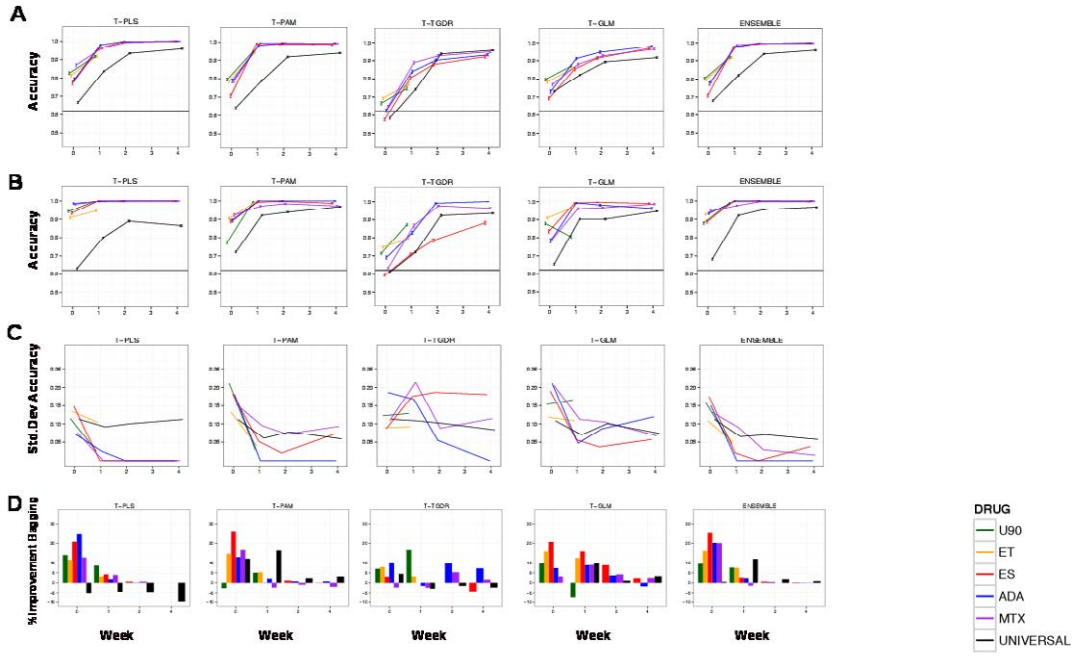

Figure 2: Performance of treatment-specific and universal classifiers on predicting PASI75 response at week 12 or later using different time spans. (A) average accuracy $\pm$ SEM over 500 bootstrapping test samples during classifiers training stage. (B) average accuracy $\pm$ SEM over 500 bootstrapping test samples after estimation of tuning parameters, weights for time-course score and bagging genes; (C) standard deviation of the accuracy as in B; (D) Percentage improvement in accuracy by coupling each algorithm with bagging.

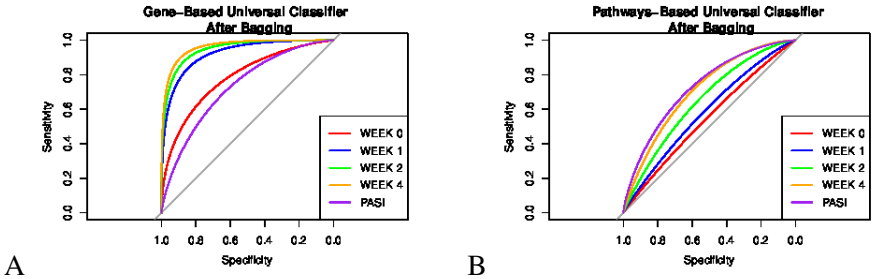

Figure 3: Receiver Operating Characteristic (ROC) curve for the final classifiers. (A) ROC curve for gene-based classifiers for each each week compared to classification that uses the PASI score evolution across time. (B) ROC curve for pathways-based classifiers for each week compared to the classification based on PASI score evolution across time.

References

[1] Katrina Abuabara, Esther E. Freeman, and Robert Dellavalle. The role of systematic reviews and meta-analysis in dermatology. *Journal of Investigative Dermatology*, 132, 2012.

[2] H. Y. Chiu, T. S.Wang, C. C. Chan, Y. P. Cheng, S. J. Lin, and T. F. Tsai. Human leucocyte antigen-cw6 as a predictor for clinical response to ustekinumab, an interleukin 12/23 blocker, in chinese patients with psoriasis: a retrospective analysis. *British Journal of Dermatology*, 2014.

[3] Joel M. Gelfand, Joy Wan, Kristina Callis Duffin, Gerald G. Krueger, Robert E. Kalb, Jamie D. Weisman, Brian R. Sperber, Michael B. Stierstorfer, Bruce A. Brod, Stephen M. Schleicher, Bruce F. Bebo Jr, Andrea B. Troxel, Daniel B. Shin, Jane M. Steinemann, Jennifer Goldfarb, Howa Yeung, and Abby S. Van Voorhees. Comparative effectiveness of commonly used systemic treatments of phototherapy for moderate to severe plaque psoriasis in the clinical practical setting. *JAMA Dermatology*, 2012.

[4] S Hanzelmann, R Castelo, and J Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC Bioinformatics*, 2013.

[5] J. Mestre-Ferrandiz, J. Sussex, and A. Towse. The r&d cost of a new medicine. Technical report, Office of Health Economics, 2012.

[6] Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9, March 2010.

[7] Mayte Suárez-Fariñas, Keyal R. Shah, Asifa S. Haider, James G. Krueger, and Michele A. Lowes. Personalized medicine in psoriasis: developing a genomic classifier to predict histological response to alefacept. *BMC Dermatology*, 2010.

[8] Suyan Tian and Mayte Suárez-Fariñas. Multi-tgdr: A regularization method for multiclass classification for microarrays experiments. *PLOS one*, 2013.

[9] Stefan Timmermans and Aaron Mauck. The promisses and pitfalls of evidence-based medicine: nonadherence to practice guidelines remains the major barrier to the successful practice of evidence-based medicine. *Health Affairs*, 2005.

[10] Federica Villanova, Paola Di Meglio, and Frank O Nestle. Biomarkers in psoriasis and psoriatic arthritis. *Annals of the Rheumatic Diseases*, 2013.

[11] Joy Wan, Katrina Abuabara, Andrea B. Troxel, Daniel B. Shin, Abby S. Van Voorhees, Bruce F. Bebo Jr., Gerald G. Krueger, Kristina Callis Duffin, and Joel M. Gelfand. Dermatologist preferences for treatments to compare in future randomized controlled comparative effectiveness trials for moderate-to-severe psoriasis. *Arch Dermatology*, 2012.