eNetXplorer: an R package for the quantitative exploration of elastic net families for generalized linear models

Julián Candia[*,†] and John S. Tsang[*,†,‡]

†*Trans-NIH Center for Human Immunology (CHI), National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA*

‡*Systems Genomics and Bioinformatics Unit, Laboratory of Immune System Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA*

E-mail: julian.candia@nih.gov; john.tsang@nih.gov

Abstract

Summary: Systems biology analysis often involves building predictive models by selecting informative features from a large number of measurements. The elastic net for generalized linear models is a popular regression and feature selection method, particularly useful when the number of features is greater than the sample size or when there exist many correlated predictor variables. This package provides a quantitative, cross-validation based toolkit to evaluate elastic net models and to uncover correlates contributing to prediction. Feature importance is evaluated by flexible criteria using out-of-bag prediction performance assessed via user-defined quality functions. Statistical significance is assigned to each model by comparison to null models generated by permutations of sample labels; analogous approaches are used to assess significance for

the contribution of individual features to prediction. This package fits linear, binomial (logistic) and multinomial models, and provides a set of standard plots, summary statistics and output tables. eNetXplorer enables quantitative, exploratory analysis to generate hypotheses on which features may be associated with biological phenotypes of interest, such as in the identification of biomarkers for therapeutic responsiveness.

Availability and implementation: The eNetXplorer R package is available under GPLv3 license at `https://CRAN.R-project.org/package=eNetXplorer`

Introduction

Rigorous, exploratory analysis for the identification of correlates and predictors in a multi-parameter/feature setting is needed in a variety of contexts, especially in systems biology where data involving a large number of features are highly prevalent. Oftentimes, bioinformatics analysis in such settings involves generalized linear regression models where observations (N) are outnumbered by parameters/features (p) measured. This class of problems can be addressed by the elastic net[1], which uses a mixing parameter α to tune the number of features used in the model continuously from ridge (more features; $\alpha = 0$) to lasso (less features; $\alpha = 1$). Algorithmically, the elastic net was efficiently implemented by coordinate descent algorithms[2,3], which, for each α, generate an entire path of solutions in the regularization parameter λ, which controls the penalty for using more parameters. While the choice of λ is usually guided by prediction performance using cross validation, α is often chosen on more subjective grounds[3].

Lasso generates parsimonious solutions involving few predictive features, particularly useful in $p \gg N$ scenarios; however, in the presence of complex correlation structures among input variables (or degeneracies), lasso will arbitrarily pick one predictor among a set of correlated features, and ignore the rest. This characteristic may lead to models that are idiosyncratic of the input dataset, as opposed to robust solutions

capturing relevant signals, or it may even lead to unstable solutions in some extreme cases. On the contrary, ridge regression promotes redundancy by shrinking correlated features towards each other, thus allowing information to be borrowed across them. In multi-parameter exploratory analysis where the primary goal is to generate hypotheses, e.g. to assess which variables correlate with a biological phenotype of interest, it is desirable to examine the entire family of elastic net models spanning the range from ridge to lasso. In this scenario, an objective, quantitative framework is needed to assess the statistical significance of individual models and, within each model, that of individual parameters/features. Towards this goal of transforming large-scale data sets into biological hypotheses, this Note describes eNetXplorer, an R-package providing a quantitative framework to explore elastic net families for generalized linear models (GLM). In the current version, three important GLM types are implemented: linear, two-class logistic, and multinomial regression. In future releases, we plan to extend it to other GLM types such as Poisson regression and the Cox model for survival data.

Workflow

Fig. 1(a-d) illustrates a typical eNetXplorer workflow to assess predictive models and parameters, here in the context of predicting antibody responses to influenza vaccination. We use data from Tsang *et al*[4], where 113 immune cell subpopulation frequencies in peripheral mononuclear cells (PBMCs) were measured in a healthy cohort before and after influenza vaccination. Included in the package is the processed dataset of cell frequencies for pre- (days -7, 0) and post-vaccination (days 1, 7, 70). Here we illustrate the use of our package by building elastic net models using data from day 7 to predict the antibody response on day 70; we know from prior analyses that day 7 contains several predictive features and overall contains the most predictive information about day 70 antibody responses (Fig. S1).

Using eNetXplorer, a family of elastic net models is generated from ridge ($\alpha = 0$) to lasso ($\alpha = 1$). For each α, the choice of λ is guided by a cross-validation grid search

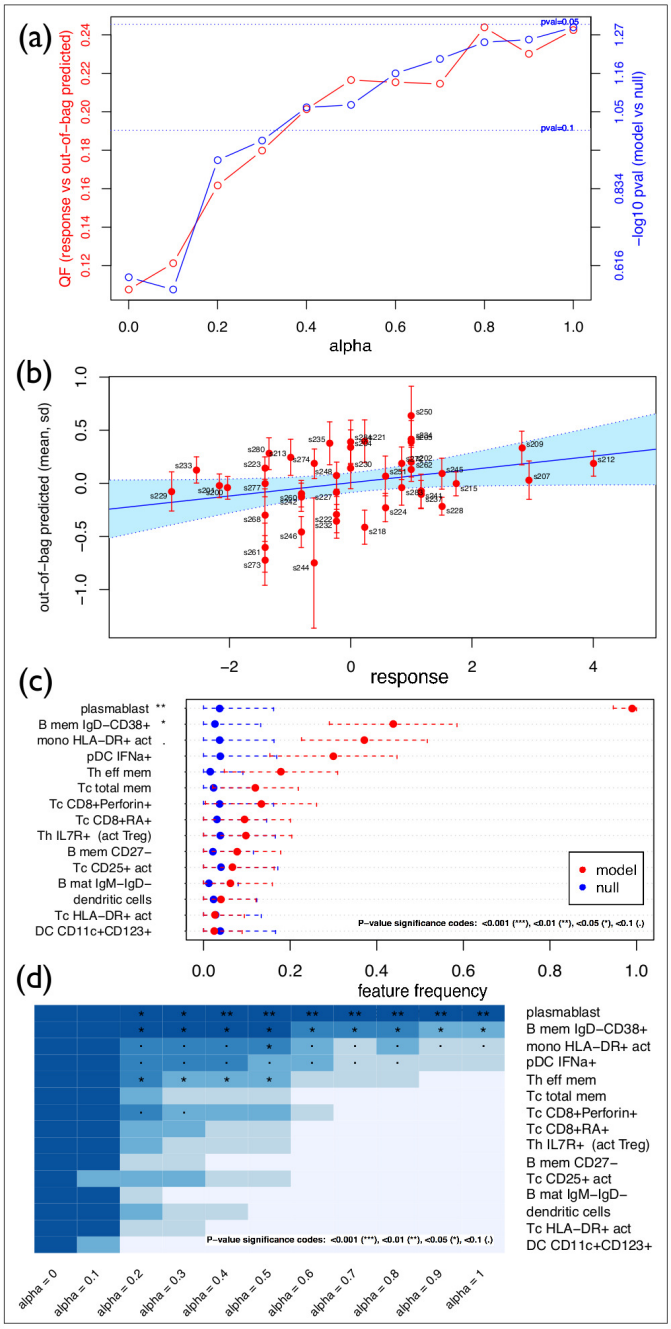

Figure 1: eNetXplorer workflow applied to a model of day 7 post-vaccination immune cell frequency predictors of day 70 H1N1 titer response. (a) Predictive performance across α: quality function and p-value of model vs null. (b) Scatterplot of true responses vs out-of-bag predictions across subjects and OOB realizations for $\alpha = 1$; best linear fit (solid line) and 95% CL (shaded region) also shown. (c) Caterpillar plot for the top features based on frequency of inclusion across cross-validation runs: model vs null for $\alpha = 1$. (d) Heatmap of feature frequencies across α.

that maximizes a quality function (QF) evaluated using out-of-bag (OOB-i.e., not used in training) instances (Fig. S2). Similarly, by evaluating QF on OOB instances, as well as the model vs null statistical significance, across the range of α, we select the best performing models in the family (Fig. 1(a)). For a model thus selected (lasso in this example), we then assess the overall quality of the OOB prediction by plotting the true vs. the predicted values (error bars indicate variability in predicted values over different random OOB realizations) and detect possible outliers (Fig. 1(b)), which may be due to biological or technical factors not taken into account in the model. The caterpillar plot in Fig. 1(c) displays the top features (here, 15 cell populations) ranked by model vs null statistical significance according to the frequency that a feature was included in the model; the top features are plasmablasts ($p < 0.01$) and IgD-CD38+ B-cell memory ($p < 0.05$). Plasmablasts at day 7 are the expected predictors of antibody responses to a vaccine challenge because they are the cells making the antibodies. Fig. 1(d) shows the same features (chosen based on the lasso solution) but here the heatmap illustrates the frequencies by which each feature was chosen to be included in each model over the entire elastic net family we explored ($\alpha = [0, 1]$). Frequencies are 1 for ridge (i.e., always included in model), but by definition none are significant compared to the null; however, as α increases, we observe the emergence of several significant features. As lasso is approached, only the top two features discussed above remain. A complementary view is offered by feature coefficient maps shown in Fig. S3, illustrating the direction and magnitude by which features contribute to the prediction of antibody responses. Taken together and with results consistent with previous analyses using different modeling approaches and discretized antibody responses[4], this example illustrates the main workflow and how our package can be used to explore building predictive models and extracting key contributors from a large number of input features.

Included in the package is also a processed dataset from a study of microRNA-based signatures of acute myeloid and lymphoblastic leukemias[5] measured in multiple cell lines and primary samples, which illustrates an eNetXplorer workflow for classification models (Figs. S4-S8).

Methods

With eNetXplorer, a family of elastic net models is generated for multiple values of α from ridge to lasso. For each α, a set of n_λ values is obtained via glmnet[2] on the full data; independently from n_λ, the user may also specify a value for $n_\lambda^{ext} > n_\lambda$ to extend the range of λ values symmetrically while keeping its density uniform in log scale. Using these values of λ, elastic net cross-validation models are generated for n_r runs, where in each run a random partitioning of samples is used for training and testing as generated by an n_f fold sampling scheme with replacement. The chosen penalization parameter λ^* is determined by maximizing a quality function (QF) that compares out-of-bag (OOB) predictions against the response (Fig. S2). User-defined QFs can be provided to the package. Otherwise, GLM-specific defaults are used: for linear regression, the default QF is Pearson's correlation (which the user can switch to Spearman's or Kendall's); for binomial models, it is accuracy; and, for multinomial models, average accuracy. For the latter two, QF defaults were chosen based on invariance under class label permutations; other popular performance measures, such as precision, recall (sensitivity), F-score, specificity, and area-under-the-curve, are not invariant[6], but may be desirable depending on the application. Individual features are characterized by their distribution of model coefficients, which we summarize by the following two measures. *Feature frequency*, ν^r_{model}, is defined as the fraction of folds (within a run) for which the feature was assigned a non-zero coefficient. Red symbols (bars) in Fig. 1(c) display the mean (standard deviation) of ν^r_{model} over all runs for a subset of selected model features. Similarly, weighted means (standard deviations) of fold-averaged non-zero feature model coefficients, κ^r_{model}, where weights are $w^r \propto \nu^r_{model}$, lead to a second type of characterization of individual features; for short, we simply call it *feature coefficient*.

A key feature of eNetXplorer is the generation of an ensemble of null models associated with each ($\alpha-$specific) member of the elastic net model family, leading to empirical statistical significance estimates for each model, as well as for individual parameters/features within each model. Each one of n_r runs are assigned into folds (based

on the same fold assignments used previously) and n_p null models per run are generated by randomly shuffling the sample label of the response; for each permutation, the overall OOB performance of the null model is evaluated via the QF, whereas the contribution of individual features is characterized by $\nu^{r,p}_{null}$ and $\kappa^{r,p}_{null}$, following analogous definitions to those given above. Blue symbols (bars) in Fig. 1(c) display the mean (standard deviation) of $\nu^{r,p}_{null}$ calculated over all run/permutation null model combinations. The empirical statistical significance of a model is hence determined as

$$\mathrm{p}_{val} = \frac{1}{1 + n_r n_p} \left\{ 1 + \sum_{r=1}^{n_r} \sum_{p=1}^{n_p} \Theta \left(QF^{r,p}_{null} - QF^{r}_{model} \right) \right\}, \tag{1}$$

where Θ is the right-continuous Heaviside step function. For sampling permutations with replacement, this expression provides a conservative estimate[7]; expressions for the exact p-value, as well as numerical approximations thereof, are provided by Phipson and Smyth[8]. The current version of eNetXplorer (1.0) implements Eq. (1). Model performance results are visualized by a summary plot, which shows the average OOB QF (red plot, left axis) and the model vs null p-value significance (blue plot, right axis) spanning the full range of α values (Fig. 1(a)). Replacing QF in Eq. (1) by ν or $|\kappa|$, our framework also provides empirical p-value estimates of the importance of individual features. Caterpillar plots are generated to display the top measures ranked by feature importance, in which significance thresholds are indicated by customary dot and asterisk annotations (Figs. 1(c) and S3(a)). Moreover, eNetXplorer also generates heatmaps of feature frequencies (Fig. 1(d)) and coefficients (Fig. S3(b)) spanning all $\alpha-$models in the elastic net family.

The same analysis strategy can be applied to any GLM in a similar fashion; two additional plot types are available to display results for binomial and multinomial models. Fig. S5(a,c) shows a graphical representation of the contingency matrix with the average number of instances predicted in each category. Fig. S5(b,d) shows boxplot representations of OOB predicted instances in each class, which is the categorical counterpart of Fig. 1(b) for linear regression. All plots shown here are standard outputs

from plotting functions included in the package; function calls include multiple graphics parameters for added display flexibility. Additional methods are available to provide summary and data export functionality to facilitate downstream analysis.

Conclusions

Uncovering correlates and predictors in a multi-parameter setting is an ubiquitous problem in systems biology. Generalized linear regression is a popular approach in this context given its flexibility, but it is often desirable to explore different levels of regularization and examine elastic net families that span the full range from ridge to lasso. By providing a quantitative, easy-to-use framework to assess the statistical significance of each model and, within each model, that of individual parameter features, eNetXplorer aims to empower users to transform large-scale data sets into biological insight.

Acknowledgements

The authors thank Angelique Biancotto, Jinguo Chen, Foo Cheung, Yuri Kotliarov, and Pamela Schwartzberg for useful discussions and feedback during the development of this package.

Funding

This research was supported by the Intramural Research Program of multiple NIH Institutes through the Trans-NIH Center for Human Immunology (CHI), NIAID, NIH.

Conflict of Interest: none declared.

References

(1) Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B*, **67**, 301-320.

(2) Friedman,J., Hastie,T. and Tibshirani,R. (2009) glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. R package available at `http://CRAN.R-project.org/package=glmnet`.

(3) Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent, *J. Stat. Softw.*, **33**, 1-22.

(4) Tsang,J.S. *et al.* (2014) Global analyses of human immune variation reveal baseline predictors of postvaccination responses, *Cell*, **157**, 499-513.

(5) Candia,J. *et al.* (2015) Uncovering low-dimensional, miR-based signatures of acute myeloid and lymphoblastic leukemias with a machine-learning-driven network approach, *Converg. Sci. Phys. Oncol.*, **1**, 025002.

(6) Sokolova,M. and Lapalme,G. (2009) A systematic analysis of performance measures for classification tasks, *Information Processing and Management*, **45**, 427-437.

(7) Ernst,M.D. (2004) Permutation methods: a basis for exact inference, *Statistical Science*, **19**, 676-685.

(8) Phipson,B. and Smyth,G.K. (2010) Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn, *Statistical Applications in Genetics and Molecular Biology*, **9**, 39.