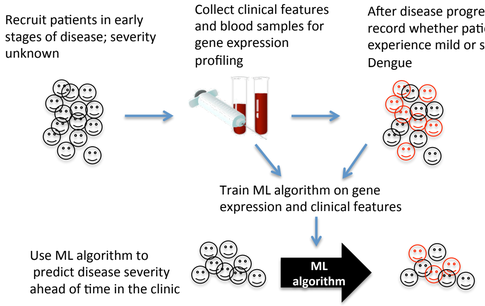
Classification of Dengue Fever Outcomes from Early Transcriptional Patterns and Clinical Features

Alec Macrae, Clement Schiano de Collela, Ellen Sebastian

Introduction

Dengue Fever is a mosquito-borne tropical disease that affects about 100 million people worldwide each year. The vast majority of Dengue infections are either nonsymptomatic or very mild, causing minor complaints such as fever, headache, and rash. However, a minority of patients - about 500,000, or 1 in 200 of all patients - develop either Dengue Hemorrhagic Fever of Dengue Shock Syndrome, which cause internal bleeding, leading to low blood pressure, and can be fatal. Given the lack of a vaccine and shortage of medical care in areas where Dengue is endemic, there is clearly an incentive to identify individuals who will likely Dengue experts have therefore tried to identify a distinct transcriptional signature, appearing in the first few days of infection, that correlates with patients' ultimate prognosis. This signature could be used to construct a low-cost diagnostic 'chip'. However, efforts at identifying such a signature have been confounded by the fact that most studies report separate sets of differentially expressed genes between mild and severe Dengue patients.

In this project, we apply machine-learning techniques to use early-infection gene expression signatures in combination with clinical data to estimate patients' likelihood of ultimately contracting severe Dengue disease. Using a Random Forest classifier in combination with a Genetic algorithm using the Akaike Information Criterion for feature selection, we achieve 95.49% 10-fold cross-validation accuracy.



Data Acquisition & Processing

Download gene expression from 6 studies of transcriptional response to Dengue. Collect gender, age, virus serotype from corresponding published papers.

Features that affect disease severity: Expression 17,609 genes, gender, age, virus serotype

Feature selection

Use as features for ML algorithm

Features unrelated to disease severity: which study, type of sample taken

Correct other features by taking residuals from ANOVA regression on study and sample type

Feature Selection Algorithms

Since we started out with 17,609 gene expression features, it was essential that we select the optimal subset of features for input to our machine learning algorithm.

Filter Feature Selection

As an exploratory early approach to feature selection, we used Filter Feature Selection to choose as features the top *n* genes most correlated with disease outcome. This approach yielded decent performance using SVM for training (Fig. 1) but did not approach the accuracy of the Genetic feature selection approach described below. 10-fold-cross validation accuracy decreased after the number of genes exceeded 50 because very few genes have a noticeable impact on disease outcomes.

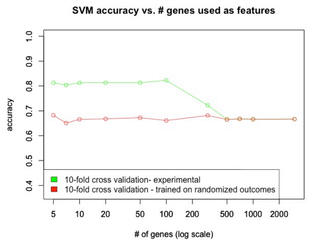


Fig. 1. SVM accuracy decreases as the number of genes, selected via Filter Feature Selection, Increases.

Genetic algorithm

Because many genes are expressed similarly, Filter Feature Selection selected a large number of redundant features. We therefore decided to try a genetic algorithm to find the best subset of genes. Using a Genetic algorithm yielded slightly better performance than Filter Feature selection using a Random Forest model for prediction, but slightly worse performance using SVM.

17,609 genes → Random Forest model (Boruta) → 100 "confirmed" or "tentative" features → Genetic Algorithm (GA) → 46 final features

Akaike Information Criterion

We think that the feature subset from Genetic algorithms was too complex, so we decided to alter the algorithm to penalize complex models. By using the Akaike Information Criterion as our fitness equation for AIC to select a model with only 9 features, we were able to achieve 95.49% 10-fold cross-validation accuracy, 3% higher than the more complex model selected by Genetic algorithms alone.

$$AIC = 2k - 2ln(L) + \frac{2k(k+1)}{n-k-1}$$

Where k = the number of features used, L = the log-likelihood of the model being tested, and n = the sample size. Since a minimum value of AIC is ideal, so higher values of *k* are penalized with a higher AIC.

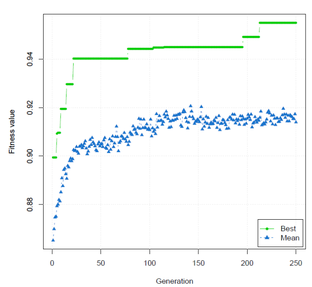


Fig. 2. accuracy of Genetic algorithm with Akaike Information Criterion, up to 250 generations (iterations) of the Genetic algorithm. The maximum accuracy obtained was 95.49%.

Machine Learning Algorithms

SVM

As an initial exploratory approach, we applied the SVM implementation implementation provided by the R package e1071. Parameters were automatically tuned using the package's tune.svm function, and we report performance resulting from optimal parameters. Ultimately, the optimal Cost parameter was 8, reflecting the fact that the data was linearly inseparable.

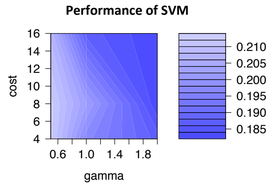


Fig. 3. SVM accuracy is optimized with high cost and low gamma.

kNN

We also experimented with k-nearest neighbors (kNN) to understand if local learning is important in classifying disease severity.

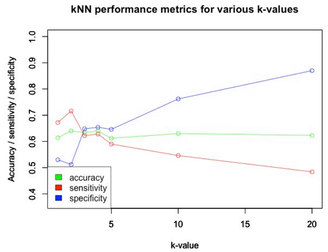


Fig. 4. Accuracy of kNN across several *k* values. We varied k-values from 1 to 20; the maximum accuracy obtained was 64.0%, for k-values between 2 and 4. However, the accuracy did not change significantly across all k values, whereas specificity and sensitivity were drastically altered by the *k* selection.

Random Forest

Finally, we applied the Random Forest implementation provided by the Caret package. We have achieved much better results with Random Forest than with any other algorithm.

Prediction \ Reference	Mild	Severe
Mild	32.7	6.2
Severe	2.9	58.1

Table 1. Confusion matrix for Random Forest prediction. This table was generated with features selected by a Genetic algorithm.

Comparison of Machine Learning and Feature Selection algorithms

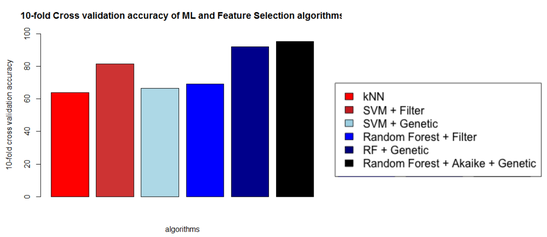


Fig. 5. Comparison of Machine Learning and Feature Selection algorithms. The highest accuracy, 95.49%, was achieved using Random Forest for prediction in combination with Genetic feature selection using the Akaike Information Criterion.

Discussion

In comparing feature selection algorithms, the Genetic algorithm is both theoretically and experimentally superior. Theoretically, it should eliminate redundant gene-expression features, therefore reducing the overall number of features. Experimentally, the Genetic algorithm is clearly superior, since it achieved 28.2% higher accuracy than FFS using Random Forest for prediction. Using the Akaike Information Criterion as our fitness metric was also effective in penalizing complex models to produce a simpler, and more accurate, set of features. In terms of training algorithms, kNN was inappropriate because local effects should not play a large role in Dengue outcomes; one patient's prognosis has no impact on another's. SVM's underperformance is due to the fact that the data is not linearly separable. Random Forest was effective because the algorithm's selection of random subsets of features helped uncover effects that would otherwise be overshadowed by more prominent features.

Reassuringly, many of the features selected as significant by our feature selection algorithm have biological implications related to an individual's ability to fight Dengue fever infection.

Feature	Gene Ontology	Comments
Primary or secondary Dengue infection	n/a	Patients suffering a secondary infection (they have been infected before) are much more prone to severe Dengue
DENND1B (DENN/ MADD domain containing 1B)	Involved in clathrin-mediated endocytosis	Endocytosis is a major activity required by immune cells to fight infection. It is often reported as enriched in gene expression studies of Dengue patients vs. healthy controls.
FBN1 (fibrillin 1)	Supports extra-cellular matrix in connective tissues	FBN1 forms microfibrils which control production of TGF-B, a growth factor important in progression to DHF.
KCNK10 (potassium channel, subfamily K, member 10)	Regulates passage of Potassium across cell membrane	Potassium transport is very generalized and affects all cells. Its role in Dengue may be related to another pathway that has not been annotated.
MYLIP (myosin regulatory light chain interacting protein)	Interacts with actin to create cell movement and growth.	Upregulation of actin-related proteins is usually observed during an immune response.
NFKBIB (nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor)	Inhibits NF-kappa-B.	NFkB triggers type-1 interferon response. Type-1 interferon response is strongly observed in transcriptional studies of Dengue.
POC1A (POC1 centriolar protein A)	Cell cycle regulation, centriole formation	Due to the high level of B cell proliferation, ontology terms related to the cell cycle are often enriched in Dengue transcriptional response studies.
CGB5 (chorionic gonadotropin, beta polypeptide 5)	Steroid hormone regulation	Steroid hormone pathways in lood are perturbed during dengue infection

Table 2. GO annotations and connections to Dengue of the optimal subset of features selected by a Genetic algorithm with Aikeke Information Criterion

Limitations

In a real-world diagnostic situation, the relative proportion of patients who will develop mild or severe Dengue would be much lower than in our study. Roughly half of our patients will go on to develop severe Dengue; in the clinic, only about 1 in 200 do. We believe that our feature selection algorithms effectively choose features that will separate the groups regardless of size; the machine learning algorithms will likely need tweaking before deployment.

Acknowledgements

We would like to thank Stephen Popper and Henry Cheng in the Relman lab at Stanford's Department of Microbiology for providing initial ideas, motivation and biological discussions, and Li Li in Atul Butte's lab in the Department of Pediatrics for providing frameworks for data acquisition and preprocessing.