CS229 Project: Classification of Dengue fever outcomes from early transcriptional patterns

Alec Macrae, Clément Schiano de Colella, Ellen Sebastian

Introduction

Dengue Fever is a mosquito-borne tropical disease that affects about 100 million people worldwide each year[1]. The vast majority of Dengue infections are either non-symptomatic or very mild, causing minor complaints such as fever, headache, and rash. However, a minority of patients - about 500,000, or 0.5% of all patients - develop either Dengue Hemorrhagic Fever of Dengue Shock Syndrome, which causes internal bleeding, leading to low blood pressure, and can be fatal[2]. Given the lack of a vaccine and shortage of medical care and medication in Dengue-endemic areas, there is clearly an incentive to identify individuals who will likely develop severe Dengue (Dengue Hemorrhagic Fever of Dengue Shock Syndrome). 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'[1]. 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 aim to 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.

Data Acquisition & Processing

Gene expression data was downloaded from the Gene Expression Omnibus[3] for six studies of the transcriptional response to Dengue fever conducted over the past seven years[1,4,5,6,7,8]. Most studies also report patients' gender, age, and virus serotype, all factors known to affect Dengue fever severity[9]. In order to standardize input across studies, genes were mapped to common identifiers using BioMart[10] and genes present in >50% of samples were considered as potential features.

Our features fell into two groups: gene expression features and non-gene expression features. We further divide the non-gene expression features into those that are useful as features, and those that have no real impact on Dengue outcome and therefore should be corrected for. The former group included gender, age, virus serotype, and whether the infection was primary (the individual's first Dengue infection) or secondary (the individual had been infected with Dengue previously). The latter group includes which study the same originated from, and what kind of blood sample was taken (whole blood or isolated white blood cells). Study origin unfortunately drove much inter-sample variation in gene expression. To correct for this, we used ANOVA regression, fitting each gene-expression trait to the study where it originated and taking residuals for further analysis.

Feature Selection Algorithms

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. Because we knew that non-gene-expression factors such as gender, age, virus serotype, and primary or secondary infection do have a large effect on Dengue outcomes, these features were included by default; FFS was limited to gene expression features. This approach yielded decent performance using SVM for training (Fig. 1) but seemed incompatible with the Random Forest algorithm, since it greatly decreased accuracy relative to the Genetic feature selection approach

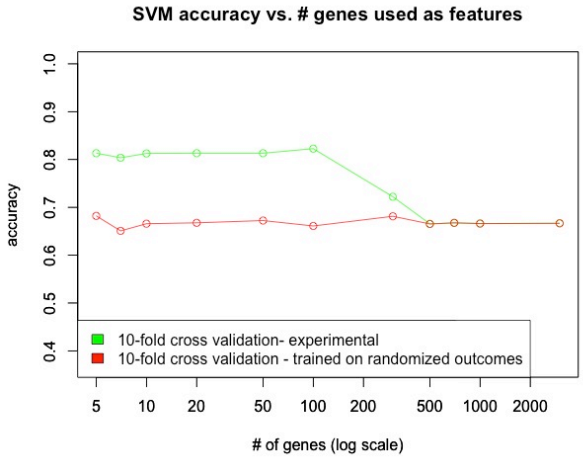


Figure 1: SVM accuracy vs. number of FFS-selected genes used as features. It is appropriate to use up to 100 genes from FFS as features; after this threshold, technical noise begins to outweigh useful information provided by the most-correlated genes.

described below.

Surprisingly, 10-fold-cross validation accuracy decreased after the number of genes exceeded 50. Using over 500 genes, accuracy was roughly the same as that produced by training the algorithm on randomized outcomes as a control. This is probably because very few genes have a noticeable impact on disease outcomes. Noise and other confounding factors outweighed biologically meaningful factors and reduced predictive power.

Backward Feature Selection

Because related genes tend to be expressed similarly, Filter Feature Selection probably resulted in an unnecessarily large number of redundant features, and therefore an unnecessarily large amount of technical noise, being incorporated into the algorithm.

In order to narrow down the number of potential feature set members, we first used the R package Boruta[11], which uses a Random Forest model to classify the features as either ”rejected” (not important), ”confirmed” (important), or ”tentative” (possibly important). This resulted in a subset of 100 ”confirmed” or ”tentative” genes. From here, our next idea was to use Backwards feature selection using the rfe function from the Caret package; but the resulting performance did not exceed 85% (figure 3). Next, we tried backwards feature selection. In this case, unlike Filter Feature Selection, Random Forest classification accuracy increased with number of genes used up to 100. However, accuracy still did not exceed 85%.

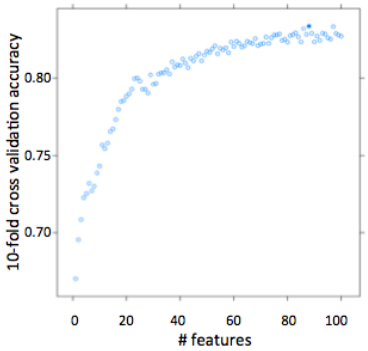


Figure 3: Backward Feature Selection

Genetic algorithm

We therefore decided to try a genetic algorithm to find the best subset of genes, using a heuristic wrapper to reduce computational expense involved in finding the optimal subset of 17,609 features. The Genetic algorithm selected 46 features as important – the smallest number of features so far – and produced Random Forest accuracy of 92%, the by far the highest accuracy yet.

Akaike Information Criterion

Although Genetic algorithms yielded much higher prediction accuracy than any other feature selection algorithm, we felt that the model selected was probably too complex, i.e., selected too many features. We noticed that 46 genes had been selected as significant, most of which had little biological connection to Dengue or immunity. Thus, we decided to tweak our model to penalize excessively complex models. The Akaike Information Criterion is a generalized measure of quality for statistical models. The equation[12] for the AIC in the case of finite sample sizes is:

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

where *k*= the number of features used, *L* = the likelihood of the model being tested, and *n* = sample size. Since a minimum value of AIC is ideal, the *k* factors act to penalize large numbers of features. We implemented AIC as a custom R function passed as the fitness parameter to GA. Whereas Genetic algorithms with GA’s default fitness function resulted in 46 genes and 92% 10-fold cross-validation accuracy with Random Forest, using AIC reduced the number of features to 9 and increased overall accuracy to 94.5%.

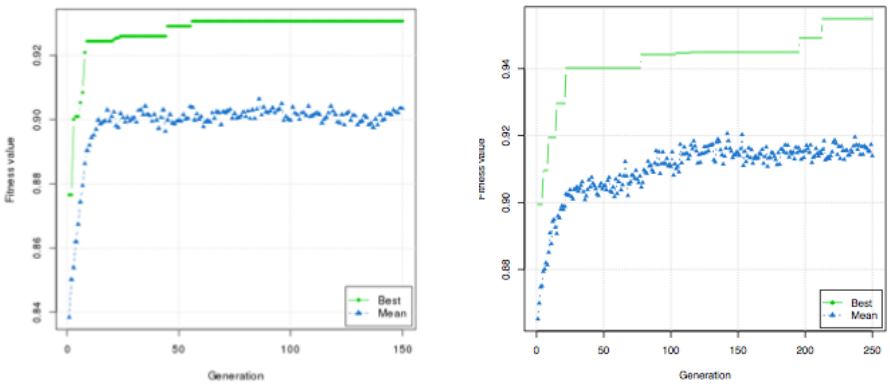


Figure 4: Genetic algorithm feature selection without (left) and with (right) Akaike Information Criterion. Using AIC reduces number of features selected from 46 to 9 and increases Random Forest accuracy from 92% to 95.5%.

Machine Learning Algorithms

SVM

As an initial exploratory approach, we applied the SVM implementation provided by the R package e1071[13]. Parameters were automatically tuned using the 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 and the algorithm needed to prioritize achieving a large margin over correctly classifying all training examples.

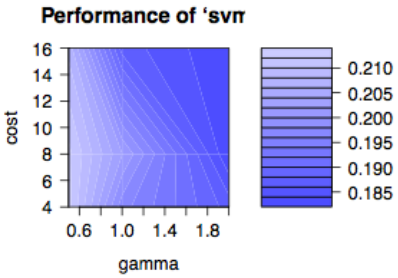
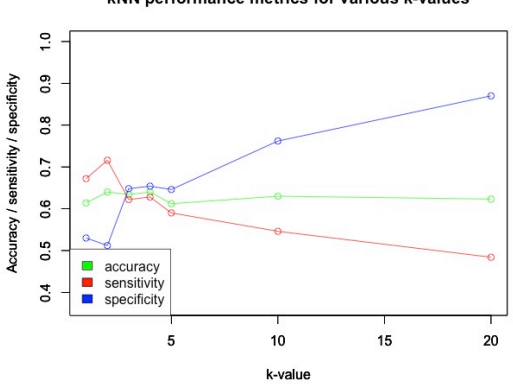


Figure 5: Tuning Cost and γ parameters for the SVM algorithm.

kNN

We also experimented with k-nearest neighbors in order to understand if local learning makes a difference in classifying disease severity. We used gene expression across samples as input features to kNN and measured accuracy using 4-fold cross-validation for varying *k*-values ranging from 1 to 20. The maximum accuracy obtained was 64.0% which is observed for *k*-values ranging from 2-4. 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. Simply by using this off-the-shelf algorithm, we have achieved much better results with Random Forest than with any other algorithm.

		Reference	
		False	True
Prediction	False	32.7	6.2
	True	2.9	58.1

Table 1: Confusion Matrix for Random Forest prediction.

Discussion

When comparing feature selection algorithms, the Genetic algorithm with Akaike Information Criterion is both theoretically and experimentally superior. Theoretically, it should eliminate redundant gene-expression features, therefore reducing the overall number of features. Indeed, even without the Akaike Information Criterion, the optimal number of genes selected by

our Genetic algorithm was 46, whereas performance was consistently low between 10 and 100 genes using FFS. Reducing the number of features is important because unimportant features will introduce technical noise and other confounding information into the algorithm. Experimentally, the Genetic algorithm is clearly superior, since it achieved 26.3% higher accuracy than FFS using Random Forest for prediction, and made little difference in accuracy using SVM for prediction. Our only concern with the Genetic algorithm without AIC was that the genes it selected appear to have little biological relevance; most genes were unannotated[14] and most of the genes that were annotated had no immunological function. In contrast, many of the top genes selected by FFS seem biologically likely to play a role in Dengue outcomes. For example, the genes PODXL2 and STAB2 were among the top features selected by Filter Feature Selection. PODXL2 plays a role in leukocyte activity, and STAB2 regulates cell adhesion and endoyctosis, two cellular activities essential to an effective primary immune response[14]. Realizing that the Genetic algorithm was probably choosing more features than actually played a significant role in Dengue outcomes, we decided to reduce the number of features the Genetic algorithm would include by incorporating the Akaike Information Criterion. In addition to increasing classification accuracy by 2.5% as mentioned, AIC also appears to have focused in on features that play a biological role in the immune response. Table 3 details the annotations of the features selected by the Genetic algorithm with Akaike Information Criterion, almost all of which had some connection to Dengue or the primary immune response. We were especially encouraged to see the demographic feature "primary vs. secondary infection" included in the list, as secondary Dengue infections are known to be much more likely to progress to Dengue Hemorrhagic Fever or Dengue Shock Syndrome as are primary infections[2].

Algorithm	Feature Selection	10-fold CV
Random Forest	Genetic + Akaike	95.49
Random Forest	Genetic	92.0 %
Random Forest	Filter Feature	69.2 %
SVM	Genetic	66.7%
SVM	Filter feature	81.5%
kNN	N/A	64.0%

Table 2: Prediction accuracy across various learning algorithms

Feature	**Gene Ontology[14]**	**Comments**
Primary vs. Secondary 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 the primary immune response 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 interac6ng protein)	Interacts with actin to create cell movement and growth	Upregulation of actin-related proteins is usually observed during an immune response due to movement and proliferation of plasmablasts.
POC1A (POC1 centriolar protein A)	Cell cycle regulation, centriole formation	Due to the high level of B cell prolifera6on, ontology terms related to the cell cycle are oden enriched in Dengue transcrip6onal response studies.
CGB5 (chorionic gonadotropin, beta polypeptide 5)	Steroid hormone regulation	Steroid hormone pathways in blood are perturbed during Dengue infection

Table 3: Gene ontology annotations for features selected by the Genetic algorithm with Akaike Information Criterion. Unlike the 46 features selected by the Genetic algorithm without AIC, most of these features have a clear connection to immunity; they may play a biological role in indicating differences in the immune response between patients who are able to effectively fight off Dengue, who do not progress to severe disease, and those with weaker immune systems, who do.

In terms of training algorithms, kNN performed the worst, Random Forest the best, and SVM in the middle. kNN was probably 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 can probably be attributed to the fact that the data does not even approach linear separability. Random Forest was probably effective because the algorithm's selection of random decision trees helped uncover effects that would otherwise be overshadowed by more prominent features.

Although we achieved good classification accuracy, several caveats must be noted. First, study-specific effects were very strong determinants of gene expression variation, which raises the possibility that much of our findings may be technical artifact. However, the fact that the features ultimately chosen were strongly associated with immunity helps assuage this fear. Second, in a real-world diagnostic situation, the relative proportion of patients who will develop severe Dengue would be much lower than in our study. As a result of study design, roughly half of our patients will go on to develop severe Dengue; in the clinic, only about 1 in 200 do. Nevertheless, we believe that our feature selection algorithms effectively choose features that will separate the groups regardless of size; but the machine learning algorithms will likely need tweaking before deployment as a diagnostic tool.

Future directions on this project may include examining the tree structure produced by the Random Forest algorithm to determine which features are most important; doing feature selection on a study-by-study basis to eliminate study-specific effects; and using bootstrap sampling to replace missing gene expression values.

References

[1] Popper SJ, Gordon A, Liu M, Balmaseda A, Harris E, Relman DA: Temporal dynamics of the transcriptional response to dengue virus infection in Nicaraguan children. PLoS Negl Trop Dis **6**(12), 1966 (2012). GEO: GSE38246

[2] Murphy, B.R., Whitehead, S.S.: Immune response to dengue virus and prospects for a vaccine. Annu. Rev. Immunol. **29**(29), 587–619 (2011)

[3] Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/

[4] Loke P, Hammond SN, Leung JM, Kim CC, Batra s, Rocha C, Balmaseda A, Harris E: Gene expression patterns of dengue virus-infected children from nicaragua reveal a distinct signature of increased metabolism. PLoS Negl Trop Dis 4(6), 710 (2010). GEO: GSE25226

[5] Simmons CP, Popper S, Dolocek C, Chau TN, Griffiths M, Dung NT, Long TH, Hoang DM, Chau NV, Thao le TT, Hien TT, Relman DA, Farrar J.: Patterns of host genome-wide gene transcript abundance in the peripheral blood of patients with acute dengue hemorrhagic fever. J Infect Dis 195(8), 1097–107 (2007). GEO: GSE40628

[6] Hoang LT, Lynn DJ, Henn M, Birren BW, Lennon NJ, Le PT, Duong KT, Nguyen TT, Mai LN, Farrar JJ, Hibberd ML, Simmons CP.: The early whole-blood transcriptional signature of dengue virus and features associated with progression to dengue shock syndrome in Vietnamese children and young adults. J Virol 84(24), 12982–94 (2010). GEO: GSE25001

[7] Long HT, Hibberd ML, Hien TT, Dung NM, Van Ngoc T, Farrar J, Wills B, Simmons CP.: Patterns of gene transcript abundance in the blood of children with severe or uncomplicated dengue highlight differences in disease evolution and host response to dengue virus infection. J Infect Dis 199(4), 537–46 (2009). GEO: GSE13052

[8] Devignot S, Sapet C, Duong V, Bergon A, Rihet P, Ong S, Lorn PT, Chroeung N, Ngeav S, Tolou HJ, Buchy P, Couissinier-Paris P.: Genome-wide expression profiling deciphers host responses altered during dengue shock syndrome and reveals the role of innate immunity in severe dengue. PLoS ONE 5(7), 11671 (2010). GEO: GSE17924

[9] Anders KL, Nguyen NM, Van Thuy NT, Hieu NT, Nguyen HL, Hong Tham NT, Thanh Ha PT, Lien le B, Vinh Chau NV, Simmons CP.: A birth cohort study of viral infections in Vietnamese infants and children: study design, methods and characteristics of the cohort. BMC Public Health. 13(937), (2013)

[10] BioMart. http://uswest.ensembl.org/biomart/martview/

[11] Kursa MB, Rudnicki WR: Feature selection with the Boruta package. Journal of Statistical Software 36(11), 1–13 (2010)

[12] Akaike Information Criterion. http://en.wikipedia.org/wiki/Akaike/

[13] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2012). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-1. http://CRAN.R-project.org/package=e1071

[14] Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S; AmiGO Hub; Web Presence Working Group. Amigo: online access to ontology and annotation data. Bioinformatics 25(2), 288–9 (2009)