Computational Intelligence applied to Human Genome Data for the Dengue Severity Prognosis

Caio Davi[1], André Pastor[2], Thiego Oliveira[3], Fernando B. Lima Neto[3], Ulisses Braga-Neto[4], Abigail W. Bigham[5], Michael Bamshad[6], Ernesto T. A. Marques[7], Bartolomeu Acioli-Santos[8]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco(IFPE)
Paulista, PE – Brazil.

[2]Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano
Serra Talhada, PE – Brazil

[3]Engenharia de Computação (eComp) – Universidade de Pernambuco (UPE)
Recife, PE – Brazil

[4]Department of Electrical and Computing Engineering – Texas A&M University
College Station, TX – USA

[5]Department of Anthropology – University of Michigan
Ann Arbor, MI – USA

[6]Division of Genetic Medicine – University of Washington
Seattle, WA – USA

[7]Department of Infectious Diseases and Microbiology,
Center for Vaccine Research – University of Pittsburgh
Pittsburgh, PA – USA

[8]Departamento de Virologia, FIOCRUZ-PE
Recife, PE – Brazil

bartacioli@cpqam.fiocruz.br

Abstract. *Dengue has become one of the most important worldwide arthropod-borne diseases around the world. Here, one hundred and two Brazilian dengue virus (DENV) III patients and controls were genotyped for 322 innate immunity gene loci. All biological data (including age, sex and genome background) were analyzed using Machine Learning techniques to discriminate tendency to severe dengue phenotype development. Our current approach produces median values for accuracy greater than 86%, with sensitivity and specificity over 98% and 51%, respectively. Genome data information from 13 key immune polymorphic SNPs was used under different dominant or recessive models. Our approach is a valuable tool for early diagnosis of the severe form of dengue infection and can be used to identify individuals at high risk of developing this form of the disease even in uninfected individuals. The model also identifies various genes involved dengue severity.*

1. Introduction

is a global public health concern that is caused by dengue virus (DENV), a positive-sense RNA virus belonging to the Flaviviridae family. It is estimated that the annual global incidence is 390 million cases, of which 96 million develop a clinically apparent self-limited disease[Bhatt et al. 2013]. In infected individuals, dengue fever (DF) occasionally progresses to dengue hemorrhagic fever (DHF) and other severe forms, that have been more recently generally classified as Severe Dengue (SD) according to the 2009 World Health Organization[World Health Organization et al. 2009] dengue classification guideline. SD classification includes several life-threatening manifestations including vascular leakage, organ failure, and shock syndrome. The mechanisms leading to the development of SD is an object of intense research. Dengue disease severity has been correlated with viral loads[Paradoa et al. 1987, Soundravally and Hoti 2007, Sakuntabhai et al. 2005], circulating viral proteins[Wang et al. 2003, Libraty et al. 2002] and exacerbated complement activity[Acioli-Santos et al. 2008, Nascimento et al. 2009].

Studies have found associations between single genetic polymorphisms (SNPs) and dengue infection phenotype in multiple genes including dendritic cell-specific inter cellular adhesion molecule 3 (ICAM-3)-grabbing nonintegrin (DC-SIGN), FCcRIIa, Transporter associated with antigen processing (TAP), Vitamin D receptor (VDR), Cytotoxic T lymphocyte-associated antigen-4 (CTLA-4), Acute plasma glycoprotein mannose binding lectin (MBL) and human platelet-specific antigens (HPA), Cytokines (IL, IFN, TNF, etc), in the Fcγ receptor IIA (a pro-inflammatory regulatory Fc receptor) gene and the vitamin D receptor and Human Leukocyte Antigen genes (HLA, i.e human histocompatibility complex)[de Carvalho et al. 2017]. These findings support the hypothesis that both adaptive memory (T-cell responses) and innate immune genes are in the dengue infection disease outcome.

There are a considerable amount of research related to dengue using computational systems[Ali et al. 2017, Muthusamy et al. 2016, Cordeiro et al. 2009]. Indeed, a lot of researches involving Machine Learning (ML) to provide differential diagnostic among DF and DHF has used a variety of techniques, such as decision trees[Tanner et al. 2008] and Support Vector Machines[Gomes et al. 2010]. But almost all of them presents limitations, such as use of clinical data and/or molecular phenotypes that are variable in time and space and/or dependent of human interpretation.

2. Material and Methods

Here we propose a novel approach to dengue infection prediction using (ML) techniques, namely SVM and ANN. This approach could be divided in the four stages described in the subsections bellow: (2.1) Data Acquisition, (2.2) Data Preprocessing, (2.3) Feature Selection, and (2.4) Patient Classification, the entire process is illustrated in Figure 1.

2.1. Data Acquisition

Patients with dengue-related symptoms were screened from three hospitals in the city of Recife, Brazil. The study was reviewed and approved by the Ethics committee of FIOCRUZ-PE: CEP/CPQAM no.11/11, C.A.A.E. 0009.0.095.000-11, IORG0001419. A set of characteristics was investigated (the type of infection, age, sex, and genetic data - 322 loci polymorphisms) over 102 patients already positively diagnosed with DF (n=27) or SD (n=75).

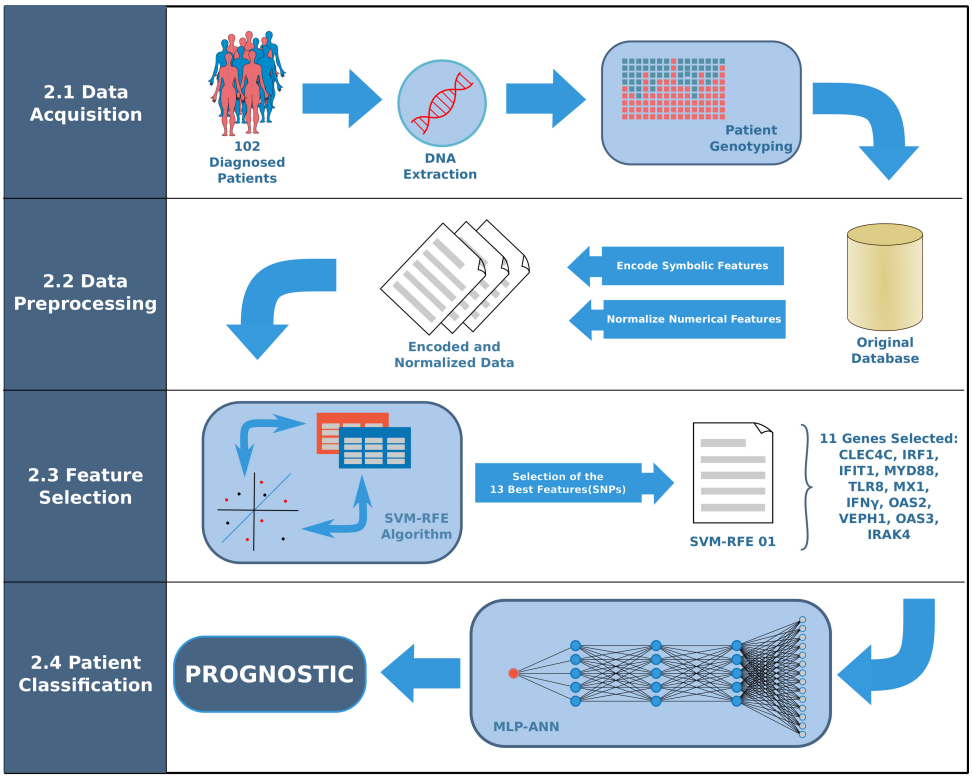

Figure 1. Flowchart of SVM-ANN/genome dengue classifier. 2.1-Data acquisition was performed by illumina genotyping of all dengue patients and then stored into a database. 2.2- Data preprocessing was performed to encode and normalize data into a suitable format for the ML step. 2.3- Feature selection was performed by keeping the best SNP subset. 2.4- A MLP-ANN classifier was learned for dengue prognosis based on the features previously selected.

2.2. Data Preprocessing

All data was encoded in values between -1 and 1. The genetic data, particularly, were encoded into indicators using a categorical scheme as homozygous dominant, heterozygous or homozygous recessive, resulting in one feature per SNP. Age, as a numeric feature, was normalized also into values between -1 and 1. Missing data was treated as a separate category. Age, the only non-categorical feature, had no missing data.

2.3. Feature Selection

Due to the high dimensionality of the data (325 categorical features and up to 900 features after converted into indicator features) and aiming to avoid the curse of dimensionality[Keogh and Mueen 2011], backward feature elimination using the SVM-RFE algorithm [Guyon et al. 2002] with a linear classification kernel[Fan et al. 2008] was performed.

The SVM-RFE model was implemented using the freely downloadable scikit-learn library provided by Pedregosa *et al.*[Pedregosa et al. 2011]. The process was repeated for all datasets using 3-fold cross-validation[McLachlan et al. 2005] to choose the SVM parameters γ and C from combinations of 0.01, 0.1, 1.0, 10.0 for each one. The process selected the value 1.0 for both. The best subset comprises 13 loci found in 11 genes, as shown in Figure 1.

2.4. Patient Classification

After that, the defined subset was used to train a Multi-Layer Perceptron-Artificial Neural Network (MLP-ANN). The MLP-ANN was implemented using the freely downloadable ML python library[Pedregosa et al. 2011]. The rectified linear unit (ReLU) function was used as the activation function and Limited-memory BFGS[Byrd et al. 1995] was used for weight optimization. The initial value of the parameter α was 0.001 and the optimal topology was found after a search in the bi-dimensional space (layers x neurons) using stratified k-fold cross-validation with k=10. The best topology found for the previously selected subset (SVM-RFE01) consisted of 3 hidden layers of 5 neurons per layer, illustrated in Figure 1 (in the MLP-ANN box).

3. Results

The Feature Selection (described in Section 2.3) found the best subset of features to classify a patient as DF ou SD. This subset comprises 13 SNPs found in 11 genes: CLEC4C, IRF1, IFIT1, MYD88, TLR8, MX1, OAS2, VEPH1, IFNγ, OAS3, IRAK4. Many of those genes have well-known influence in the immune system and the metabolic pathways of each one are subject of further studies in our research.

Those genes are used as input for a MLP-ANN. The accuracy estimation for this subset of features were obtained by the bolstered resubstitution method[Braga-Neto and Dougherty 2004a]. The variance of the bolstering kernels were set using the "Naïve Bayes" method[Jiang and Braga-Neto 2014]. Bolstered estimation is more accurate than cross-validation for small datasets[Braga-Neto and Dougherty 2004b]. The estimated accuracy rates are reported in Table 1. Also displayed, for comparison, are the stratified 10-fold cross-validated accuracy rates obtained in the network selection step (as can be seen, these accuracy rates are inflated by selection bias).

Table 1. Estimated statistics for our two best classifiers.

Subset	Cross-Validation	Bolstered Resubstitution	
SVM-RFE01	96%	86.1%	Accuracy
	100%	98.64%	Sensitivity
	85%	51.85%	Specificity

4. Conclusion

Here we applied ML techniques, namely SVM and ANN, to develop a classifier based on genomic polymorphism to predict the risk of SD. In the Feature Selection step (see Figure 1) we have used a SVM to select the best subset to be used in this classification. This subset consists in 13 SNPs located in 11 innate immune genes: CLEC4C, IRF1, IFIT1, MYD88, TLR8, MX1, OAS2, VEPH1, IFNγ, OAS3, IRAK4. The role of these genes in the immune mechanisms involved in the severe dengue phenotype are very promising, though currently this research is in a preliminary phase. The use of genome data to predict diseases has several advantages, especially due it can be done at any time and in a broad human sample tissue, during early virus infection and/or before the infection itself.

The training dataset used for training the SVM/ANN were well characterized. It was shaped by data from 13 SNPs to produce a classifier with accuracy level greater than

86%, with a sensitivity of 98,64%, and specificity around 51%. This type of diagnostic tool is useful for patient triage, especially during disease outbreaks.

The method presented here provided very robust results for a prognosis (for dengue severity) classifier, as demonstrated by the error assessment calculations. It is able to select the optimal loci combination data and, then, to correctly (pre) classify the patient that will develop severe phenotype based only in its genome background. Our method can be easily replicate for other genetic based/genetic influenced diseases, helping to find optimal loci sets to understand the molecular architecture of different pathologies, and driving to potential therapeutics target genes.

References

Acioli-Santos, B., Segat, L., Dhalia, R., Brito, C. A., Braga-Neto, U. M., Marques, E. T., and Crovella, S. (2008). Mbl2 gene polymorphisms protect against development of thrombocytopenia associated with severe dengue phenotype. *Human immunology*, 69(2):122–128.

Ali, I., Humayun, F., Azam, S., Munir, A., Rizwan, M., et al. (2017). Computational tool for classification of dengue virus. *J Appl Bioinforma Comput Biol 6*, 3:2.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., et al. (2013). The global distribution and burden of dengue. *Nature*, 496(7446):504.

Braga-Neto, U. and Dougherty, E. (2004a). Bolstered error estimation. *Pattern Recognition*, 37(6):1267–1281.

Braga-Neto, U. M. and Dougherty, E. R. (2004b). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Cordeiro, M. T., Braga-Neto, U., Nogueira, R. M. R., and Marques Jr, E. T. (2009). Reliable classifier to differentiate primary and secondary acute dengue infection based on igg elisa. *PloS one*, 4(4):e4945.

de Carvalho, C. X., Cardoso, C. C., Kehdy, F. d. S. G., Pacheco, A. G., and Moraes, M. O. (2017). Host genetics and dengue fever. *Infection, Genetics and Evolution*.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Gomes, A. L. V., Wee, L. J., Khan, A. M., Gil, L. H., Marques Jr, E. T., Calzavara-Silva, C. E., and Tan, T. W. (2010). Classification of dengue fever patients based on gene expression data using support vector machines. *PloS one*, 5(6):e11267.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Jiang, X. and Braga-Neto, U. (2014). A naive-bayes approach to bolstered error estimation in high-dimensional spaces. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, pages 1398–1401. IEEE.

Keogh, E. and Mueen, A. (2011). Curse of dimensionality. In *Encyclopedia of Machine Learning*, pages 257–258. Springer.

Libraty, D. H., Endy, T. P., Houng, H.-S. H., Green, S., Kalayanarooj, S., Suntayakorn, S., Chansiriwongs, W., Vaughn, D. W., Nisalak, A., Ennis, F. A., et al. (2002). Differing influences of virus burden and immune activation on disease severity in secondary dengue-3 virus infections. *The Journal of infectious diseases*, 185(9):1213–1221.

McLachlan, G., Do, K.-A., and Ambroise, C. (2005). *Analyzing microarray gene expression data*, volume 422. John Wiley & Sons.

Muthusamy, K., Gopinath, K., and Nandhini, D. (2016). Computational prediction of immunodominant antigenic regions & potential protective epitopes for dengue vaccination. *The Indian journal of medical research*, 144(4):587.

Nascimento, E. J., Silva, A. M., Cordeiro, M. T., Brito, C. A., Gil, L. H., Braga-Neto, U., and Marques, E. T. (2009). Alternative complement pathway deregulation is correlated with dengue severity. *PloS one*, 4(8):e6782.

Paradoa, M. P., Trujillo, Y., and Basanta, P. (1987). Association of dengue hemorrhagic fever with the hla system. *Haematologia*, 20(2):83–87.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Sakuntabhai, A., Turbpaiboon, C., Casadémont, I., Chuansumrit, A., Lowhnoo, T., Kajaste-Rudnitski, A., Kalayanarooj, S. M., Tangnararatchakit, K., Tangthawornchaikul, N., Vasanawathana, S., et al. (2005). A variant in the cd209 promoter is associated with severity of dengue disease. *Nature genetics*, 37(5):507.

Soundravally, R. and Hoti, S. (2007). Immunopathogenesis of dengue hemorrhagic fever and shock syndrome: role of tap and hpa gene polymorphism. *Human immunology*, 68(12):973–979.

Tanner, L., Schreiber, M., Low, J. G., Ong, A., Tolfvenstam, T., Lai, Y. L., Ng, L. C., Leo, Y. S., Puong, L. T., Vasudevan, S. G., et al. (2008). Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS neglected tropical diseases*, 2(3):e196.

Wang, W.-K., Chao, D.-Y., Kao, C.-L., Wu, H.-C., Liu, Y.-C., Li, C.-M., Lin, S.-C., Ho, S.-T., Huang, J.-H., and King, C.-C. (2003). High levels of plasma dengue viral load during defervescence in patients with dengue hemorrhagic fever: implications for pathogenesis. *Virology*, 305(2):330–338.

World Health Organization, W., for Research, S. P., in Tropical Diseases, T., of Control of Neglected Tropical Diseases, W. H. O. D., Epidemic, W. H. O., and Alert, P. (2009). *Dengue: guidelines for diagnosis, treatment, prevention and control*. World Health Organization.