African Journal of Biotechnology Vol. 12(23), pp. 3724-3730, 5 June, 2013
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB12.1860
ISSN 1684-5315 ©2013 Academic Journals

Full Length Research Paper

Applying machine learning to predict patient-specific current CD_4 cell count in order to determine the progression of human immunodeficiency virus (HIV) infection

Yashik Singh[1,2]*, Nitesh Narsai[2] and Maurice Mars[1]

[1]Department of Telehealth, Nelson R. Mandela School of Medicine, Durban, South Africa.
[2]School of Information Systems and Technology, University of KwaZulu Natal, Durban, South Africa.

Accepted 10 October, 2012

This work shows the application of machine learning to predict current CD_4 cell count of an HIV-positive patient using genome sequences, viral load and time. A regression model predicting actual CD_4 cell counts and a classification model predicting if a patient's CD_4 cell count is less than 200 was built using a support vector machine and neural network. The most accurate regression and classification model took as input the viral load, time, and genome and produced a correlation of co-efficient of 0.9 and an accuracy of 95%, respectively, proving that a CD_4 cell count measure may be accurately predicted using machine learning on genotype, viral load and time.

Key words: Human immunodeficiency virus (HIV), antigens, CD4, computational biology, artificial intelligence, data mining, pattern recognition.

INTRODUCTION

The current trend in patient healthcare is personalized medicine where treatment is individualized, rather than a response to set physical presentations. Thus, access and interpretation of personal patient information is vital, in order to provide a sustainable and useful medical service. The science of information systems, management and interpretation plays an important role in the continuity of care of patients. This is becoming more evident in the treatment of HIV/AIDS. In 2008, there were between 30 and 36 million HIV-positive patients around the world. HIV/AIDS is the leading cause of death in sub-Saharan Africa (Campbell et al., 2008) and is currently the fastest growing epidemic in South Africa (Simon-Meyer et al., 2002). Southern Africa continues to have the largest burden of HIV, with sub-Saharan Africa home to 67% of HIV-positive patients. The South African Department of Health listed HIV/AIDS as being one of the top four health priorities in the country (Campbell et al., 2008). It is estimated that 50% of all new infections in sub-Saharan Africa are from South Africa (Giarelli et al., 2000), currently with 5.5 million confirmed cases of the disease (Rispel et al., 2009).

HIV infection can be effectively managed with antiretroviral (ARV) drugs, but close monitoring of the disease progression is vital. Monitoring of the progression of the disease is made even more important due to the emergence of HIV drug resistance, especially in developing countries with limited resource. HIV drug resistance refers to the inability of the ARV drug to reduce the viral reproduction rate sufficiently. Poor management of HIV drug resistance will lead to opportunistic infections that make treatment of HIV more difficult and even may lead to fatalities. The Health Systems Trust reported that almost 44% of all HIV positive patient deaths are due to AIDS defined conditions, which includes HIV drug resistance and poor monitoring of virus progression (Health Systems Trust, 2011). Given enough time, drug resistance is inevitable due to selective pressure and the high

*Corresponding author. E-mail: singhyashik@gmail.com.

mutation rates of HIV (Vercauteren and Vandamme, 2006).

Laboratory (Fahey et al., 1990; Moss et al., 1998) and clinical (Cahn et al., 1991; Montaner et al., 1992) markers of disease progression can monitor HIV infection. Common laboratory tests include viral load (measurement of HIV nucleic acid concentration), CD_4 lymphocyte counts or CD_4/CD_8 ratio (quantitative measurement of particular lymphocyte cells that indicate the strength of the immune system), HIV p24 antigenaemia (p24 antigen count increases one to three weeks after initial infection and is present before HIV antibodies are formed), erythrocyte sedimentation rate (a non-specific marker of inflammation), and serum-β-2-microglobulin because high concentrations are a good predictor of AIDS (Morfeldt-maringnson et al., 2002). Common clinical markers of disease progression are weight loss, mucocutaneous manifestations, bacterial infections, chronic fever, chronic diarrhoea, herpes zoster, oral candidiasis, and pulmonary tuberculosis (Morgan et al., 2002).

One of the best available surrogate markers for HIV progression is the use of CD_4 cell count information (Post et al., 1996; Schechter et al., 1994). In developed countries, and areas where there is a low rate of HIV infection, CD_4 cell count is recognized as a standard measure of immunodeficiency in HIV positive patients (Mwamburi et al., 2005). Although this is also standard of care in developing countries, the measurement of CD_4 cell count requires many complex and expensive flow cytometric procedures which burden the minimal resources available (Schechter et al., 1994). There have been previous attempts to predict CD_4 cell count information using cheaper chemical assays and even correlating a patient's total lymphocyte count (TLC) with CD_4 cell counts using logistic and linear regression (Schechter et al., 1994; Mwamburi et al., 2005).

In developing countries, TLC may be used to make treatment decisions when CD_4 cell count is not available and patients are symptomatic (Daka et al., 2008). Machine learning has been used as a prediction tool in many areas of medicine, including predicting ARV drugs. HIV positive patients are resistant [(Vercauteren and Vandamme, 2006) MRI segmentation for breast images] (Kannan et al., 2011) to platelet transfusion requirements for leukemia (Ho and Chang, 2011). Machine learning may also be used to predict CD_4 cell count information. Machine learning is an artificial intelligence computer science technique that tries to find a mathematical model that map between inputs and outputs of a domain problem. There are two stages of using machine learning techniques. These are: creating the mathematical model by learning mappings between given input and output, secondly using the model to predict an output given unseen input.

Wang et al. (2002) developed a prototype neural network machine-learning algorithm on a small dataset, and predicted the viral load of HIV-positive patients from genogenotype and treatment data with 75% accuracy. Larder et al. (2007) later created a genotype and treatment history based input neural network model that predicted viral load with 69% accuracy. Altmann et al. (2008) created a machine-learning algorithm that predicts a dichotomized virological response that is, success or failure of therapy, with 80% success. This was later changed by predicting the probability of treatment success (Altmann et al., 2009) based on a degree of predicted HIV drug resistance.

These previous studies were based on predicting viral load or a dichotomized response. These are valuable in the treatment of HIV, however, most treatment guidelines in developing counties are based on CD_4 guided management of HIV. CD_4 guided management of HIV has been described as well-tolerated and cost-saving (Ananworanich et al., 2005). CD_4 guided treatment is a proven approach to long term management of HIV (Dyner et al., 2006).

There has, however, been no computer model developed to predict current CD_4 cell count. Singh and Mars (2010) describes a model that predicts how CD_4 count may change in the future and also how ARV drugs may affect future changes in CD_4 cell count. The model however, does not predict current CD_4 count, and thus does not speak to the advantages of knowing how to allocate current recourse and how to manage patients at the current point in time.

The aim of this study was to apply machine-learning techniques to produce a mathematical model that can predict a measure of CD_4 cell count at an individual HIV-1 positive patient level that is, an actual CD_4 cell count or whether a patient's CD_4 cell count is less than 200. This amounts to using machine learning to create a regression and classification mathematical model, respectively.

MATERIALS AND METHODS

Dataset

Separate patient datasets containing protease (PR) or reverse transcriptase (RT) genome sequences, CD_4 cell counts or viral loads were obtained from the Stanford HIV drug resistance database (http://hivdb.stanford.edu/). These de-identified datasets are publically available.

Pre-processing

PR genome sequences, CD_4 cell count, viral load and the number of weeks from the baseline measure of CD_4 cell count for each patient sample was determined by joining individual datasets using sample identifier (unique number that identifies a sample) and date. This was also done for the RT dataset, creating two datasets to apply the machine learning techniques on. The PR population dataset consisted of approximately 4,500 data elements, while the RT population dataset contained approximately 2,500 sequences. Each protease sequence consisted of 99 amino acids from position 1 to 99, and each reverse transcriptase sequence comprised of 201 amino acids from position 40 to 240. The sequences were processed such that at any position, 1 represented an amino acid

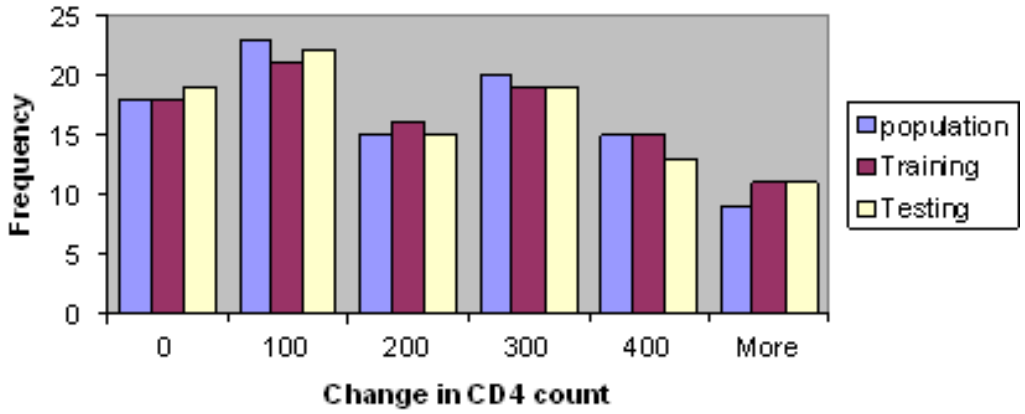

Figure 1. The distribution of the PR population, test and training datasets.

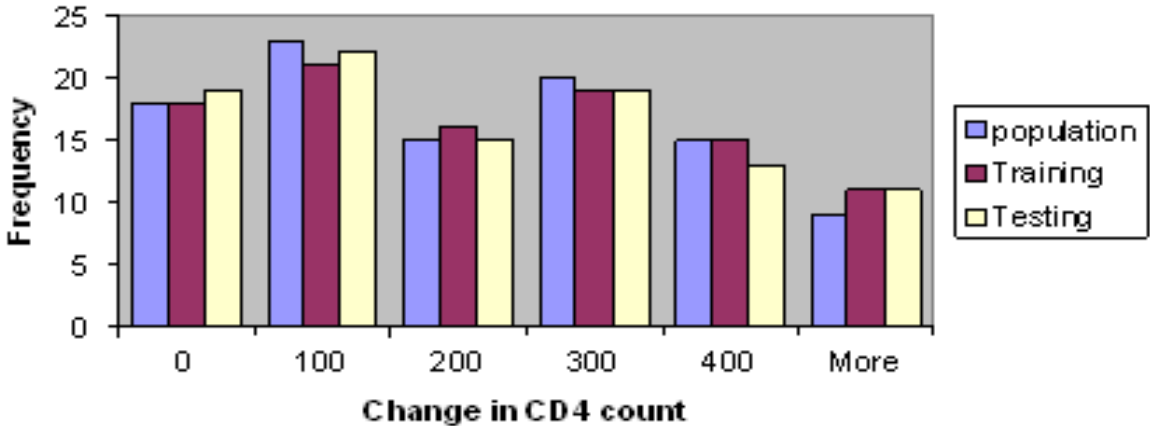

Figure 2. The distribution of the RT population, test and training datasets.

mutation and a 0 represented no mutation. The PR and RT datasets were further processed to reduce the number of amino acid positions that would be used for learning and testing. The number of mutations that occurred in the entire dataset at each position was calculated. All positions were the total number of mutations was less than 5% were removed. This resulted in the PR dataset having genome sequence data in 31 positions, while the RT dataset had data in 47 positions. 200 random data elements were removed from each dataset and formed a testing set. Training was done on the remaining data elements of the PR and RT datasets. The distributions of the change in CD_4 cell counts found in the test, training, and PR and RT population datasets are shown in Figures 1 and 2. Since the distributions of the test, training and population datasets are similar, we can conclude that the training and testing datasets are representative of the RT and PR population datasets.

Input of the machine learning algorithms

Three different groups of inputs were created and each were fed into the machine learning algorithms separately, forming six models for regression and six for classification. These input groups were: input 1, consisted only of viral load; input 2, consisted of viral load and genome sequence; input 3, consisted of viral load, genome sequence and number of weeks the CD_4 cell count was taken from baseline CD_4 cell count and; input 4, consisted of genome sequence only.

Output of the machine learning algorithms

The output of the regression machine learning models was the actual CD_4 cell count. For classification, the data elements were grouped into two categories, as shown in Equation 1.

$$classification = \begin{cases} 1, CD_4 < 200 \\ 0, CD_4 \geq 200 \end{cases} \quad (1)$$

Machine learning techniques

Two machine learning techniques were used to find a mapping between the three input groups and the measure of CD_4 cell count. The first was a statistical technique called support vector machines (SVMs) and the other was a neural network (NN) which is a technique based on the structure of the human brain's neuron. SVMs were chosen due to their ability to learn high dimension inputs and they are a robust learning method (Daka et al., 2008; Navia-Vazquez, 2007; Bugers, 1998). Neural networks (Draghici et al., 2003) are one of the most common machine learning algorithms that have proven to be valuable in many domain environments.

Support vector machines

SVMs work by embedding data into a higher dimensional vector

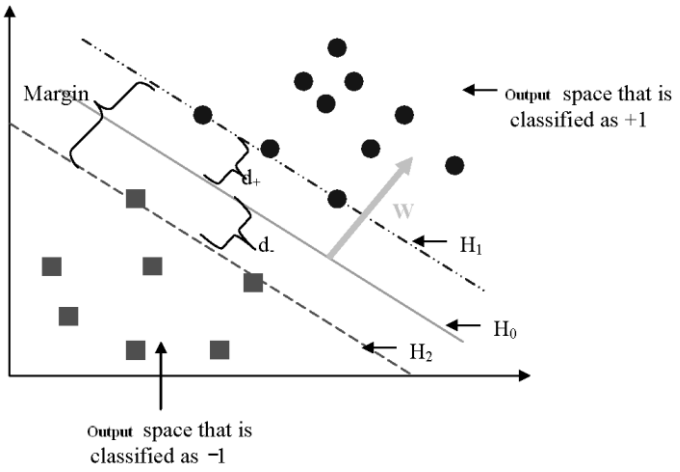

Figure 3. Graphical representation of SVMs.

space and then attempts to find linear relations in that space. SVMs have been described as having the properties of duality, ability to incorporate kernels, margin maximization, convexity, and sparseness. The simplest possible SVM is one that learns data in linear space. This means that there must exist a linear separating hyperplane (H0) that completely separates the input space from its output space. This separating boundary divides the input space in such a way that all input space elements that lie on one side of the boundary have the output space value of +1, while those on the opposite side have the value -1. This separation is shown in Figure 3.

SVMs try to maximize the margin between (H0) and (H1), that is ||W||. After incorporating the Wolf dual, Lagrangian multipliers and kernels (K(), which is used to convert a non-linear search space into a linear one by increasing dimensionality), SVMs reduces to solving Equation 2.

$$L_{dual} = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^{N} \alpha_i \alpha_j \gamma_i \gamma_j \mathrm{K}(X_i, \mathrm{X}_j)$$

$$W = \sum_{i=1}^{N} \alpha_i \gamma_i X_i$$

$$\sum_{i=1}^{N} \alpha_i \gamma_i = 0$$

$$0 \leq \alpha_i \leq \zeta \qquad (2)$$

Where, Imp is a user defined value that assigns the importance of an error, external stimuli called the bias (β), α_i are Lagrangian multipliers, X_i is the input space, and y_i is the output space for i = 1 … number of data elements.

SVMs solve regression by converting it into a classification problem. Assume that a SVM is trying to approximate a function y = f(x) with z = g(x). Thus, for any given x, if regression (f(x),g(x)) = 0 then it is considered to be classified correctly (Equation 3).

$$\text{Regression}(f(x),g(x)) = \begin{cases} 0, & \|f(x)-g(x)\| < \varepsilon \\ \|f(x)-g(x)\| - \varepsilon, & \text{otherwise} \end{cases} \qquad (3)$$

LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used to create the SVM model in this study. A radial base function kernel and a cost of 3000 were used. A cost of 3000 was chosen as it gave the highest accuracy using a trial and error approach in preliminary training.

Neural networks

All neural networks contain an electronic representation of the human neuron. These networks use and interconnect these neurons in various manners in order to perform learning. Similar to the dentrites attached to the axon of the human neuron, the electronic neuron has channels which transport input (α) into the core of the neuron. The electronic neuron has weights (ϖ) and an external stimuli called the bias (β). The electronic neurons may or may not fire, depending on the value (θ) of the product and summation of the inputs, weights and bias, and on the type of activation function (Ψ) that is used. Mathematically, a neuron can be described as shown in Equations 4 and 5. This study used a single layer, recurrent neural network with 8 neurons which was created using NeuroSolutions (http://www.neurosolutions.com/).

$$\theta = \sum_{i=0}^{num\ of\ inputs} \alpha \bar{\omega} + \beta \qquad (4)$$

$$\Psi(\theta) = \begin{Bmatrix} \text{Fire} \\ \text{Dont Fire} \end{Bmatrix} \qquad (5)$$

Statistical analysis

Pearson's correlation (r) was used to measure how closely the regression model correlates to known CD_4 counts. The correlation was computed as shown in Equation 6.

$$r = \frac{\sum_{i-1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1) S_x S_y} \qquad (6)$$

Where, n is the number of tuples in a dataset, X_i is the i^{th} desired CD_4 cell count, Y_i is the i^{th} predicted CD_4 cell count, S_x and S_y are the standard deviations of the desired and predicted CD_4 counts.

Accuracy, sensitivity, specificity, positive predictive value and negative predictive value were used to determine how close the classification model predicted the known CD_4 cell counts. Accuracy was defined as the percent of correct predictions. Sensitivity was calculated as the number of predicted patients whose CD_4 cell counts were greater and equal to 200, compared to all patients whose CD_4 cell counts were greater and equal to 200. Specificity was determined by calculating the number of predicted patients whose CD_4 cell counts were less than 200 compared to all patients whose CD_4 cell counts were less than 200. Positive predictive value is defined as the proportion of patients with CD_4 cell counts less than 200 who are correctly diagnosed, similarly for negative predictive value. The best regression mathematical model was determined by comparing coefficient of correlations of all regression models using the Fisher-Z score. Multivariate random block analysis of variance (RBD) was used to compare accuracy, sensitivity, specificity, positive predictive values and negative predictive values in order to determine the optimal classifier.

RESULTS

Accuracies, sensitivities, specificities, positive predictive

Table 1. Accuracy and regression results produced by support vector machines and neural networks.

Dataset	Type	Support vector machine						Neural network					
		Classification/%					Reg\r	Classification/%					Reg\r
		Acc	Sen	Spec	PPV	NPV		Acc	Sen	Spec	PPV	NPV	
PR	Input 1	61	64	73	64	73	0.51	78	28	84	64	54	0.41
	Input 2	89	89	89	87	91	0.76	78	31	72	53	51	0.19
	Input 3	87	81	91	86	88	0.77	75	42	73	61	56	0.19
	Input 4	71	70	71	59	80	0.6	79	40	81	68	58	0.42
RT	Input 1	74	60	79	59	79	0.52	80	91	77	57	91	0.43
	Input 2	91	86	94	89	92	0.90	88	48	84	75	62	0.51
	Input 3	95	94	92	86	97	0.85	83	63	77	73	68	0.39
	Input 4	81	78	82	63	90	0.55	80	46	82	72	60	0.46

Acc = accuracy; Sen = sensitivity; PPV = positive predictive value; NPV = negative predictive value; Reg = regression.

Table 2. Results of RBD analysis of variance ([a]P > 0.05, [b]P < 0.05, [c]P < 0.001).

Test	F_{input}	$F_{machine}$
Pillai's Trace	2.031[b]	11.844[c]
Wilks' Lambda	4.124[c]	48.881[c]
Hotelling's Trace	6.604[c]	180.696[c]
Roy's Largest Root	39.796[c]	436.766[c]

Table 3. Post Hoc testing for differences in input and machine learning algorithms in terms of various measures ([a]P > 0.05; [b]P < 0.05; [c]P < 0.001).

Measure	F_{input}	$F_{machine}$
Accuracy	52.780[c]	692.248[c]
Sensitivity	14.189[c]	120.832[c]
Specificity	35.078[c]	633.867[c]
PPV	29.759[c]	304.117[c]
NPV	24.700[c]	274.955[c]

value (PPV), nearest point problem (NPP) and the coefficient of correlation for the support vector machine and neural network models are shown in Table 1. Input 1 produced an average accuracy of 73%; input 2, an average accuracy of 86%; input 3, 85% and input 4, an average accuracy of 77%. The highest accuracies produced with SVMs were 95% for RT and nput 3, while neural networks achieved 88% with input 2 and RT. Multivariate random block design analysis of variance (RBD) was used to determine if the differences shown in Table 1 are statistically significant, taking into account the effects of the variance on the datasets, input groups and machine learning techniques.

Random block design is a statistical theoretical framework that is used to analyze variance. It is similar to a two factor fixed-fixed design, but is applied to datasets were there is only a single value for each factor. RBD was performed with machine learning techniques as the treatment factor, input groups, protease and reverse transcriptase genome datasets as blocks, and the accuracies, sensitivities, specificity, positive predictive value and negative predictive value as the dependent variables. These results are shown in Table 2.

Duncan post hoc testing on the effect of the different machine learning algorithms on accuracy, specificity, PPV and NPV are shown in Table 3. Similarly, Duncan post hoc results for input groups are shown in Table 4 (1 to 2 refers to comparing input 1 to input 2, 1 to 3 is comparing input 1 to input 3 and 2 to 3 comparing input 2 to Input 3). Table 5 contains the results of the Fisher-Z scores for the comparison of the regression models.

DISCUSSION

The $F_{machine}$ and p values in Table 2 are less than the critical value at 95% confidence level for all of Pillai's Trace, Wilks Lambda, Hoteling's Trace and Roys's Largest Root. Thus, there is a statistically significant difference between the two machine learning algorithms. Duncan post hoc tests (Table 3) indicate that there are statistical differences between the learning methods for accuracy, sensitivity, specificity, PPV and NPV. There is an 11% difference in accuracies between SVM's and NN's, taking into account all input groups. The higher overall accuracies, sensitivities, specificities, PPN and NPV across all input groups as a whole indicate that SVM's outperform NN's.

The fact that SVM's outperformed NN's confirms the ability of SVM's to better learn highly dimensional, complex and non-separable data. Thus, when the input space dimensionality increased by adding genome data, SVM's prediction improved. NN's, on the other hand, does not learn high dimensional data well, thus learning was not improved much after adding genome data. F_{input} and p values, in Table 3 indicate that there is a statistically

Table 4. Ad Hoc to determine differences in input groups.

Input	Accuracy	Sensitivity	Specificity	PPV	NPV
1-2	0.044	0.042	0.034	0.003	0.007
1-3	0.036	0.042	0.034	0.003	0.006
2-3	0.986	1.000	1.00	0.672	0.973
1-4	0.008	0.009	0.007	0.002	0.009
2-4	0.030	0.009	0.050	0.040	0.009
3-4	0.040	0.009	0.050	0.040	0.008

Table 5. Fisher's Z score for comparing regression ($^{a}P > 0.05$, $^{b}P < 0.05$, $^{c}P < 0.001$).

Dataset	Input	Fisher's Z
PR	Input 1	-1.97^{b}
	Input 2	-13.96^{c}
	Input 3	-14.78^{c}
	Input 4	-4.92^{c}
RT	Input 1	-1.85^{a}
	Input 2	-22.84^{c}
	Input 3	-19.23^{c}
	Input 4	-4.68^{c}

significant difference between the input groups. Post hoc testing (Table 4) indicates that, in terms of accuracy, all input groups were statistically different, except between input 2 and input 3. The same results are seen in sensitivity, specificity and PPV.

Taking into account accuracy, sensitivity, specificity, PPV, and NPV input 1 performed the worst. Input 4 produced a better prediction model input space. These results indicate the importance of genome data in the prediction model. The prediction results were improved by using input 2 and input 3 (that is, adding viral load). This is due to the fact that the viral load is indicative of the ARV drugs being administered. In order to produce a good prediction model, it is vital that the current ARV treatment is taken into account.

There is no statistically significant difference between input 2 and input 3. The number of weeks from the treatment baseline does not seem to improve classification. This may be due to the fact that the effect of the ARV over the time duration is reflected by the mutations in the genome. P values in Table 5, indicates that there is a significant statistical difference between the ability of SVM's to predict CD_4 cell count than NNs (except for RT input 1). This again, proves the superior ability of SVM to learn highly complex data.

The previous studies discussed, predicted treatment outcomes in terms of viral load, total lymphocyte count, or success or failure of treatment. However CD_4 guided treatment important in the management of HIV (Dyner et al., 2006) is even recognized as the gold standard for treatment (Obirikorang et al., 2012). This study contributes by allowing the prediction of current CD_4 cell count that may be used when managing HIV through CD_4 cell count guided treatment. It has been suggested that CD4 guided treatment should be included in the decision making process of complex HIV cases where there is clinical benefit with failing treatment but with immunologic benefits where CD4 count > 350 cells/mm^3 (Dyner et al., 2006). Thus, the ability of this tool to predict current CD_4 cell count will be invaluable when one needs constant monitoring of the patients CD_4 cell count.

This tool is also beneficial in terms of resource allocation although standard of care that continuously monitor CD_4 cell counts are resource intensive. Instead, one may predict CD_4 cell count over a longer period of time with a single genome. Only if the algorithm predicts a low CD_4 count, will the patient be sent for a laboratory CD_4 cell count test, thus reducing the burden on resources.

Conclusion

Results obtained from this study indicate that a measurement of CD_4 cell count can be successfully predicted using machine learning. Actual CD_4 cell counts and predicting if a patient's CD_4 cell count is less than 200 is possible using protease and reverse transcriptase genomes, viral load and number of weeks from baseline measure. Support vector machines were shown to out perform neural networks, both in predicting CD_4 cell count and if the CD_4 cell counts are less than 200.

Genome data were also shown to improve the prediction abilities of the machine learning algorithms.

To the knowledge of the authors, predicting CD_4 cell counts using protease and reverse transcriptase genomes, viral load and number of weeks from baseline measure is novel. This study is the first objective of a project trying to predict a patient's future CD_4 cell count trend when there is a treatment change. Future work will include replacing the viral load measurement in the predictive model with actual therapy information and adding more classification output categories. This will aid physicians in guiding HIV antiretroviral treatment.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health Fogarty International Centre (grant number 5D43TW007004-11).

REFERENCES

Altmann A, Rosen-Zvi M, Prosperi M, Aharoni E, Neuvirth H et al (2008). Comparison of Classifier Fusion Methods for Predicting Response to Anti HIV-1 Therapy. PLoS ONE. 3(10): e3470.

Ananworanich J, Siangphoe U, Hill A, Cardiello P, Apateerapong P, Hirschel B, Mahanontharit A, Ubolyam S, Cooper D, Phanuphak D, Ruxrungtham K (2005). Highly active antiretroviral therapy (HAART) retreatment in patients on CD4-guided therapy achieved similar virologic suppression compared with patients on continuous HAART: the HIV Netherlands Australia Thailand Research Collaboration 001.4 study. J Acquir Immune Defic Syndr 39(5): 523-529. Bugers CJC (1998). A tutorial on support vector machines for pattern recognition. Data mining and Knowledge discovery. 2:121-67.

Daka D, Loha E (2008). Relationship between Total Lymphocyte count (TLC) and CD4 count among peoples living with HIV, Southern Ethiopia: a retrospective evaluation. AIDS Res Ther. 5:26.

Dyner T, Cafaro V, Boly L (2006). CD4 guided treatment manage-ment.: AIDS 2006 - XVI International AIDS Conference: Abstract no. THPE0149

Larder B, Wang D, Revell A, Montaner J, Harrigan R, De Wolf F et al. (2007) The development of artificial neural networks to predict virological response to combination HIV therapy. Antivir Ther. 12(1):15-24.

Mwamburi DM, Ghosh M, Fauntleroy J, Gorbach SL, Wanke CA (2005). Predicting CD4 count using total lymphocyte count: A sustainable tool for clinical decisions during HAART use. Am J Trop Med Hyg. 73(1):58-62.

Navia-Vazquez Z (2007). Support vector perception. Neurcomputing. 70:1089–95.

Post FA, Wood R, Maartens G (1996). CD4 and total lymphocyte counts as predictors of HIV disease progression. Q J Med. 89:505-8.

Schechter M, Zajdenverg R, Machado LL, Pinto ME, Lima LA, Perez MA (1994). Predicting CD4 Counts in HIV-infected Brazilian Individuals: A Model Based on the World Health Organization Staging System. J Acquir Immune Defic Syndr. 7(2):163-8.

Wang D, DeGruttola V, Hammer S, Harrigan R, Larder B, Wegner S et al. (2002). A Collaborative HIV Resistance Response Database Initiative: Predicting Virological Response Using Neural Network Models. Poster presentation at: The XI International HIV Drug Resistance Workshop. Seville.

Altmann A, Däumer M, Beerenwinkel N, Peres Y, Schülter E, Büch J et al (2009). Predicting the Response to Combination Antiretroviral Therapy: Retrospective Validation of geno2pheno-THEO on a Large Clinical Database. JID. 199:999-1006.

Cahn P, Perez H, Casiro A, Crinberg N, Muchinik G (1991).Progression of HIV-disease: the Buenos Aires cohort study. Int conf AIDS. 8:157 (Abstract no. PuC 8029).

Campbell C, Nair Y, Maimane S, Sibiya Z (2008). Supporting people with aids and their careers in rural South Africa: Possibilities and challenges. Health Place. 14:507-18.

Fahey JL, Taylor JM, Detels R, Hofmann B, Melmed R, Nishanian P et al (1990). The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus type 1. N Engl J Med. 322:166-72.

Giarelli E, Jacabs AL (2000). HIV/AIDS care in Kwazulu-Natal, South Africa: An interview with Dr. Leana Uys. JANAC. 11(4): 52-67.

Health Systems Trust (2011). Percentage of deaths due to AIDS Health Indicators.

Statistics South Africa: Statistical release P0302 Mid-year estimates 2011. from http://indicators.hst.org.za/indicators/StatsSA/.

Ho W, Chang C (2011). Genetic-algorithm-based artificial neural network modeling for platelet transfusion requirements on acute myeloblastic leukemia patients. Expert. Syst. Appl. 38(5):6319-6323

Kannan SR, Ramathilagam S, Devi R, Sathya A (2011). Robust kernel FCM in segmentation of breast medical images. Expert Syst. Appl. 38(4):4382-4389.

Montaner JSG, Le TN, Le N, Craib KJP (1992). Schechter MT. Application of the World Health Organization system for HIV infection in a cohort of homosexual men in developing a prognostically meaningful staging system. AIDS. 6:719-24.

Morfeldt-maringnson L, Boumlttiger B, Nilsson B, von Stedingk L (2002). Clinical signs and laboratory markers in predicting progression to AIDS in HIV-1 infected patients. Scand. J. Infect. Dis. 23(4):443-9.

Morgan D, Mahe C, Mayanja B, Whitworth JAG, Kilmarx PH (2002). Progression to symptomatic disease in people infected with HIV-1 in rural Uganda: prospective cohort study. BMJ. 324:193-7.

Moss AR, Bacchetti P, Osmond D, Krampf W, Chaisson RE, Stites D et al (1998). Seropositivity for HIV and the development of AIDS and AIDS related condition: three year follow up of the San Francisco general hospital cohort. Br. Med. J. 296:745-50.

Obirikorang C, Quaye L, Acheampong I (2012). Total lymphocyte count as a surrogate marker for CD4 count in resource-limited settings. BMC Infect Dis. 12(1):128.

Draghici S, Potter RB (2003) Predicting hiv drug resistance with neural networks. Bioinfor. 19(1):98-107.

Rispel CL, Metcalf AC (2009). Breaking the silence: South African HIV policies and the needs of men who have sex with men. Reprod. Health. 33:133-42.

Simon-Meyer J, Odallo D (2002). Greater involvement of people living with HIV/AIDS in South Africa. Eval Prog Plan. 25:3850–55.

Singh Y and Mars M (2010). Support vector machines to forecast changes in CD4 count of HIV-1 positive patients. Sci Res Essay 5(17):2384-2390.

Vercauteren J, Vandamme AM (2006). Algorithms for the interpretation of HIV-1 genotypic drug resistance information. Antivi Res **7**1(2–3): 335-342.