SHORT-TERM HIV THERAPY RESPONSE PREDICTION USING SEQUENCE INFORMATION

Cem Meydan and Uğur Sezerman

Biological Sciences and Bioengineering, Sabancı University
Sabancı University, Tuzla, 34956, Istanbul, Turkey
phone: + (90) 216 483 9513, fax: + (90) 216 483 9550, email: ugur@sabanciuniv.edu

ABSTRACT

HIV causes 3 million deaths annually. Advancements in medical sciences have enabled us to manage the infection with drug therapies in the recent years. However, HIV-1 has high viral variability and is likely to evolve resistance against these drugs, considering a high correlation between the evolutionary rate and disease progression. It is important to understand the genetic blueprint of the virus and the marker mutations that are linked with disease progression under treatment.

For a data set containing clinical patient data at the beginning of the therapy, and the reverse transcriptase (RT) and protease (PR) nucleotide sequences of HIV-1 virus, we developed an algorithm to extract a number of features and predict the short term progression of the disease with response to the therapy and find the important positions within the sequences. The algorithm resulted in around 30 positions that can predict the disease progression with AUC of 0.824 and accuracy of 0.737, better than the standard methods and comparable to the best methods available on such a data.

1 INTRODUCTION

Human immunodeficiency virus (HIV) has caused the deaths of approximately 25 million people between its discovery in 1981 to 2006, and further causes 2.4-3.3 million deaths annually [1]. Even though newly discovered drug therapies have increased the life expectancy in HIV+ people, the virus is highly variable and develops mutations that provides resistance against the drugs in time [2]. The replication of the virus is highly error prone (around 1 viral mutation occurring every replication cycle), and due to this high evolution rate, the virus exists within a patient as a complex mixture of related but distinguishable variants called quasispecies [2]. High evolutionary rate of some HIV-1 variants within a patient have been correlated with disease progression since the selection pressure eliminates non-resistant variants [3].

For this reason, it is important to find the combination of mutations that infer this resistance to specific drugs. These mutations may able us to develop new treatments more effectively. Moreover, depending on the viral RNA levels in the plasma and the variants of the virus present in a patient, the treatment options may change since some drug susceptibility of HIV variants are different [2] . Therefore it is important to be able to predict disease progression and response to antiretroviral therapy beforehand using the pre-treatment clinical data and the genetic blueprint of the virus.

Many computational studies have been conducted on HIV, to determine important sequence positions, to predict drug resistance of a variant, to predict disease progression with time. For estimating drug specific genotypic susceptibility scores (GSS) to predict virologic response, algorithms such as HIVdb [4], ViroSeq [5], ANRS [6] and Rega [7] have been developed. These algorithms can be used to predict drug susceptibility of a variant given its RT and protease sequences and they can identify mutations that create resistance or susceptibility to specific drugs [8]. These four algorithms have been shown to predict the response with area under the receiver operator characteristic curve (AUC) value of 0.76, and their weighted combination increased the AUC to 0.80 on a data set to predict response during treatment change episodes [9].

For this study, we used clinical information and some of the sequence information of the virus present to predict whether the viral load decreases 100-fold within a 16-week time frame, and tried to pinpoint the markers that affect the outcome of the treatment.

2 METHODS

2.1 Data Set

Data set is curated by William Dampier and obtained from Kaggle [10]. It consist of samples from 1612 patients infected with HIV-1 virus, with each sample containing the pre-treatment values of viral load (viral particles in 1mL of blood, log10 scale) and CD4+ cell count (in 1 ml of blood), plus the nucleotide sequence of Reverse Transcriptase (RT) and Protease (PR) enzymes of the virus. The class label is a binary value indicating whether the patient has shown response to drug therapy after 16 weeks, with response being defined as 100-fold decrease in viral load. The data combines samples with different drug therapies due to the variability of therapeutic intervention. Due to the drug cocktails not being universal, patients were treated with different combinations of 13 drugs, 1 to 3 at a time [10]. However the data set was prepared to question the existence of markers that indicate good/bad progression independent of the therapy chosen. In addition, the treatment data is omitted in the original dataset due to difficulty of representing different cocktails with different (and unknown) dosages. For these reasons, the treatment information is not used.

2.2 Feature Extraction and Selection

The data contains both clinical data in the format of numerical values and sequences. To use the sequence information we need to extract features to formats usable by the machine learning algorithms. For that purpose, the RT and the PR sequences are aligned within themselves using ClustalW [11] for both classes. After alignment, each nucleotide in the aligned RT and PR sequence of a sample is expanded into a binary vector of size 5 (A, C, G, T, and gap). If the nucleotide in the current position of the current row is A, the vector is created as 1 for the "A" and 0 for the remaining four elements. The ambiguous characters that define more than one nucleotide (e.g. R for purine, Y for pyrimidine, N for any nucleotide and others as well) are also converted accordingly using IUPAC nucleotide code definitions.

Another feature that is commonly used in sequence datasets is the frequency of all k-mers in the sequences. We created features for all possible 3-mers and added their frequencies for all of the samples.

The features up to now include only the position/mutation information in the sequences and local k-mer frequencies. To capture further information that may exist in a combination of different positions, a hidden Markov model (HMM) of the sequences are built using the software HMMER [12]. The profile HMM is built for both the positive and the negative class using the alignment of the sequences (by aligning each class within themselves), and each sequence is tested against both of those profiles, resulting in a score and an e-value for both classes and both RT and PR sequences, for a total of 8 features.

Apart from the mutation positions that discriminate between two classes and the global HMM profiles, we also looked at the possibility of more generalized local sequence motifs in the RT and PR sequences. We used MEME (Multiple EM for Motif Elicitation) motif discovery suite [13] and DEME algorithm [14] for discriminative motif discovery. Discriminative motifs are sequence motifs that differentiate the two sets of sequences, and are found by searching for patterns that are overrepresented in one class and underrepresented in the other. We searched for overrepresented sequences in both positive and negative sets using the training set. The discovered motifs for both classes were searched in the test set and the search score with the respective software is taken as the motif score. However, due to high mutability of HIV, the found motifs were either not consistent in the class they are supposed to be in, or occurred both in positive and negative classes with very high probability. Due to the low information gain, they were removed by the feature selection algorithms in nearly all of the runs.

The above steps results in more than 9,000 separate sequence features, majority from the expansion of gene sequences. Many of those features are unimportant and carry no information with regards to the class value; this reduces the classification accuracy significantly. Also, to find the important mutations and positions that affect the therapy response, we need a way to filter out the unimportant features. For these reasons, application of feature extraction methods was necessary.

We used Evomarker [15, 16], a genetic algorithm based biomarker detection method, to find a near optimal subset of features. This resulted in 7 features. However, Evomarker assumes the resulting feature set will be used as biomarkers and tries to minimize the feature count aggressively. This may remove some important features if they are correlated with other features in the subset. Since we want to find the important positions as well as obtain high prediction accuracy, we also added the selected features in an SVM based sequential floating forward selection (SFFS) and the top 100 features when ranked by their information gain [17]. We created multiple feature sets with feature count ranging from 7 to 180 using the combination of these 3 methods. Of those, results of 2 are presented here, one with 32 features and one with 112 features. The other sets fared nearly the same in terms of accuracy.

To see the importance of each feature and a visual representation of their relationship with samples, we created a 2D linear projection [18] of the data using 5 features. Those 5 features were selected from all of the features excluding sequence position data by running the VizRank heuristic [19] for 2000 generations. The rotation of the axes and the final projection was optimized using the FreeViz algorithm [20] to optimize separation of data points. The result is given in Figure 1. The viral load before the treatment has been consistently selected in all of the feature sets and we can see that it is the most single important feature for separation of two classes. Interestingly, the other clinical information, CD4+ cell count, has not been selected in neither of the feature sets and was not used in the projection even if added manually.

Another point is that higher viral load at the start of treatment is significantly correlated with better therapy response (can also be seen in Figure 1), which is directly the opposite of what one would expect.

Sequence positions were not added to the pool of features during the course of the VizRank algorithm due to their count and the computational complexity of the selection. To gather the informative sites in sequences, we calculated the information gain and the gain ratio [17] of each position separately. The results for both RT and PR sequences are given in Figure 2. On average, we can see some single peaks (very specific informative sites, such as position 559 in RT sequences) and regions with higher average information count (from positions 250 to 350 in RT sequences). It is important to note that the given position data is from the complete multiple alignment, and the numbers will change between different sequences due to the presence of gaps in the complete alignment.

2.3 Classification

For classification, the continuous features such as viral load, CD4 cell count, e-values from HMMER results are normalized between 0 and 1. The resulting data is classified using SVM (support vector machines) and random forest. Random forest classification is performed under Orange [21]. SVM classification is performed under Weka using LibSVM [22-24]. For SVM kernel, radial basis function is used with cost parameter of 200. Since the data set is unbalanced, the positive class was weighted 4 times the negative class, taken as the inverse square of the ratio of 1:2 positive to negative samples.

This imbalance in the data set also affects the performance estimates. Since there are roughly twice the number of negative samples, classifying all samples as negative yields an accuracy of 66%, which is an overestimate. To correct for this bias, we used balanced accuracy in our tests, given as:

$$\text{Balanced Accuracy} = 1 - \frac{\frac{\text{FP}}{\text{N}} + \frac{\text{FN}}{\text{P}}}{2}$$

where FP and FN are the count of, respectively, false positives and false negatives, N and P are the count of all negative and positive samples.

If the classifier has equal discriminatory power for both classes, this term becomes equal to the normal accuracy. However if the accuracy is high just because the classifier takes advantage of the imbalanced data set, balanced accuracy will reduce and normalize the accuracy [25]. By this normalization, classification of an unbalanced set (with respect to number of positive and negative samples) will always give the baseline accuracy of 0.5 if all of the samples are classified as the majority class, instead of the probability of the class with the higher prior. Using the balanced accuracy was shown to be a better estimate of performance [26]. This balanced accuracy measure was used as the metric to optimize in the Evomarker genetic algorithm for feature selection. This ensures that the genetic algorithm weighs both classes the same and does not try to select features by their accuracy dominantly on the negative class.

3 RESULTS

Results of the classification with two different feature sets are given in Table 1. Our results are from 5-fold cross validation, whereas the comparison results are trained and tested on separate sets.

AUC (area under the ROC curve) value is a better estimate of classifier performance than accuracy. The classification with highest AUC, SVM using 112 features, has AUC of 0.824, which means that a randomly selected positive sample will get a higher score in prediction than a randomly selected negative sample with 82.4% probability. Even though SVM using 32 features has less AUC and lower conventional accuracy, it has the highest balanced accuracy with 0.737, which means that this prediction method would fare better if we weigh both of the classes the same in the loss function and not proportional to the sample size in the class.

Although AUC and accuracy values are quite high, which means a good portion of the data from both classes are separated cleanly, specificity values are low enough to cause problems. The threshold separating the classes can be shifted with weighing one class more than the other; however either specificity or sensitivity will suffer in any case. Even though the errors in the positive class are weighted more than the errors in the negative class, the data unbalance still makes it harder to separate both classes at once. Since the classifiers we use are greatly affected by unbalanced class priors, we evaluated our method on a balanced set. The balanced set was created by randomly undersampling the negative class to create 1:1 ratio of positive to negative examples and repeating the 5-fold cross validation. As we can see in Table 1, this significantly increases specificity without sacrificing much sensitivity, giving a more balanced prediction for both classes. Although the ratio of positive examples to negative examples may not be near 1:1 in real world examples, using bootstrapping with undersampling to achieve uniform classes will increase the accuracy.

For comparison, the top results from the Kaggle competition are also shown [10]. Note that these results are from a separate training and test set (training with 1000 samples, testing with 692 samples), and may not be directly comparable. A problem in the competition was that the training set was not representative of the test set and the selection was not random (e.g. “Patient id”/row number was correlated with the response, class priors were significantly different in training and testing sets etc.). Due to these reasons we opted to use 5-fold cross validation instead of testing the method on a separate test set. In addition, since the alignments and feature selection of our method were done in a cross validation scheme, using those features in the test set would have introduced some selection bias into our results. But for completeness, the (possibly slightly biased) test results for SVM has an AUC of 0.734 and accuracy of 0.727. With these results, our method performs higher than the maximum accuracy in literature, but is surpassed by the competition winner in terms of accuracy (AUC values unknown). It is important to note that the competition winner uses data segmentation to combine samples into groups and classifies them separately, in a way performing ensemble classification [10]. As mentioned, the data is from a collection of patients treated with different drug cocktails and it is possible that different groups overlap partially/completely with those treatment types. It is known that variants have different susceptibility to different drugs and there are methods that exploit this information to predict which treatment should be used to maximize treatment response. We created a heat map of the distance matrix of samples using Euclidian distance between their feature vectors to show whether there were any obvious groupings. As we can see in Figure 3, the samples can be clustered into specific units with low intra-class distance and higher inter-class distance. Our method can be improved by training different classifiers for different clusters to achieve greater accuracy. However there is no way to know if those clusters overlap with the treatment type since that information is missing.

4 DISCUSSION

It is possible to predict HIV progression in patients using clinical data and the sequence markers from virus variants up to acceptable accuracy, even in the absence of treatment information. Although the accuracy can be increased using multiple time points and more clinical information, the data easily obtainable from plasma at one time point is still useful and much easier to obtain and track. Apart from the prediction step, we were able to determine important positions that are useful for class separation, though checking the validity of those mutations and annotating them requires expert review to be useful for patient treatment. When found, such markers are useful for selecting antiretroviral drugs to increase drug susceptibility for a specific variant present in the patient.

It should be noted that aligning these sequences is quite hard as a result of the high mutability of HIV-1. Since the alignment is crucial in the latter stages, this presents a serious problem. While there are expert curated alignments of sequences obtained from specific clinical trials, there were none containing all of our sequences, and the ambiguity in the final alignment is high, even for curated data. For future work, using machine learning methods that can work with unaligned sequences can be necessary. An extension to SVM method is the string kernel [27]. The string kernel works directly with strings without the need for an alignment [28]. Another alignment-free method is to use semi-supervised learning on a network created by the relative complexity measure as the distance measure. The class labels can be propagated through the links using the network topology and the distances. This step was removed from our algorithm due to time limitations; however this method is very suitable to the context of the problem and may provide better predictions.

5 REFERENCES

[1] Joint United Nations Programme on HIV/AIDS, “Overview of the global AIDS epidemic,” *2006 Report on the global AIDS epidemic*, 2006.

[2] S. Y. Rhee, M. J. Gonzales, R. Kantor *et al.*, “Human immunodeficiency virus reverse transcriptase and protease sequence database,” *Nucleic Acids Research,* vol. 31, no. 1, pp. 298-303, Jan 1, 2003.

[3] K. Kozaczynska, M. Cornelissen, P. Reiss *et al.*, “HIV-1 sequence evolution in vivo after superinfection with three viral strains,” *Retrovirology,* vol. 4, pp. -, Aug 23, 2007.

[4] T. F. Liu, and R. W. Shafer, “Web resources for HIV type 1 genotypic-resistance test interpretation,” *Clin Infect Dis,* vol. 42, no. 11, pp. 1608-18, Jun 1, 2006.

[5] S. H. Eshleman, J. Hackett, Jr., P. Swanson *et al.*, “Performance of the Celera Diagnostics ViroSeq HIV-1 Genotyping System for sequence-based analysis of diverse human immunodeficiency virus type 1 strains,” *J Clin Microbiol,* vol. 42, no. 6, pp. 2711-7, Jun, 2004.

[6] Agence National de Recerche sur le SIDA (ANRS), “ANRS genotypic resistance guidelines (version 13),” no. Accessed 3 June 2009, 2009.

[7] K. Van Laethem, A. De Luca, A. Antinori *et al.*, “A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients,” *Antivir Ther,* vol. 7, no. 2, pp. 123-9, Jun, 2002.

[8] D. Frentz, C. A. Boucher, M. Assel *et al.*, “Comparison of HIV-1 genotypic resistance test interpretation systems in predicting virological outcomes over time,” *PLoS One,* vol. 5, no. 7, pp. e11505, 2010.

[9] S. Y. Rhee, W. J. Fessel, T. F. Liu *et al.*, “Predictive value of HIV-1 genotypic resistance test interpretation algorithms,” *J Infect Dis,* vol. 200, no. 3, pp. 453-63, Aug 1, 2009.

[10] A. Goldbloom, “Kaggle, a platform for data prediction competitions,” 2009.

[11] J. D. Thompson, T. J. Gibson, and D. G. Higgins, “Multiple sequence alignment using ClustalW and ClustalX,” *Curr Protoc Bioinformatics,* vol. Chapter 2, pp. Unit 2 3, Aug, 2002.

[12] R. Durbin, *Biological sequence analysis : probabalistic models of proteins and nucleic acids*, Cambridge, UK New York: Cambridge University Press, 1998.

[13] T. Bailey, and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." pp. 28-36, 2004.

[14] E. Redhead, and T. L. Bailey, “Discriminative motif discovery in DNA and protein sequences using the DEME algorithm,” *BMC Bioinformatics,* vol. 8, pp. -, Oct 15, 2007.

[15] A. Küçükural, R. Yeniterzi, S. Yeniterzi *et al.*, “Evolutionary selection of minimum number of features for classification of gene expression data using genetic algorithms,” *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, 2007.

[16] C. Meydan, and O. U. Sezerman, “Biomarker discovery for toxicity,” *Neurocomputing,* vol. 73, no. 13-15, pp. 2384-2393, Aug, 2010.

[17] S. Kullback, “The Kullback-Leibler Distance,” *American Statistician,* vol. 41, no. 4, pp. 340-340, Nov, 1987.

[18] Y. Koren, and L. Carmel, “Visualization of labeled data using linear transformations,” *Infovis 2002: Ieee Symposium on Information Visualization 2003, Proceedings*, pp. 121-128, 248, 2003.

[19] G. Leban, B. Zupan, G. Vidmar *et al.*, “VizRank: Data visualization guided by machine learning,” *Data Mining and Knowledge Discovery,* vol. 13, no. 2, pp. 119-136, Sep, 2006.

[20] J. Demsar, G. Leban, and B. Zupan, “FreeViz--an intelligent multivariate visualization approach to explorative analysis of biomedical data,” *J Biomed Inform,* vol. 40, no. 6, pp. 661-71, Dec, 2007.

[21] J. Demsar, B. Zupan, G. Leban *et al.*, “Orange: From experimental machine learning to interactive data mining,” *Knowledge Discovery in Databases: Pkdd 2004, Proceedings,* vol. 3202, pp. 537-539, 2004.

[22] C.-C. Chang, and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001.

[23] M. Hall, E. Frank, G. Holmes *et al.*, “The WEKA Data Mining Software: An Update,” *SIGKDD Explorations,* vol. 11, no. 1, 2009.

[24] Y. EL-Manzalawy, and V. Honavar, "WLSVM: Integrating LibSVM into Weka Environment," 2005.

[25] K. H. Brodersen, C. S. Ong, K. E. Stephan *et al.*, “The balanced accuracy and its posterior distribution,” *Proceedings of the 20th International Conference on Pattern Recognition*, pp. 3121-3124, 2010.

[26] D. R. Velez, B. C. White, A. A. Motsinger *et al.*, “A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction,” *Genetic Epidemiology,* vol. 31, no. 4, pp. 306-315, May, 2007.

[27] J. Shawe-Taylor, and N. Cristianini, *Kernel methods for pattern analysis*, Cambridge, UK ; New York: Cambridge University Press, 2004.

[28] S. Boisvert, M. Marchand, F. Laviolette *et al.*, “HIV-1 coreceptor usage prediction without multiple alignments: an application of string kernels,” *Retrovirology,* vol. 5, pp. -, Dec 4, 2008.

	AUC	Accuracy	Balanced Acc.	Precision	Sensitivity	Specificity
Our method, Feature Set A (1692 samples, 112 features)						
SVM	0.824	0.745	0.660	0.757	0.912	0.407
Random Forest	0.800	0.727	0.632	0.740	0.913	0.351
LAD Tree	0.769	0.719	0.666	0.773	0.822	0.510
Our method, Feature Set B (1692 samples, 32 features)						
SVM	0.736	0.679	0.737	0.919	0.927	0.547
Random Forest	0.778	0.731	0.659	0.767	0.867	0.450
LAD Tree	0.775	0.738	0.683	0.781	0.845	0.520
Our method, Balanced set (32 features, 1062 samples)						
SVM	0.800	0.728	0.728	0.738	0.707	0.749
Random Forest	0.762	0.685	0.685	0.685	0.685	0.685
SVM on data w/o feature selection (8990 features)	0.668	0.656	0.612	0.744	0.888	0.448
Comparison (on a separate test set, 692 samples)						
Kaggle, literature max.	-	0.708	0.708	-	-	-
Kaggle, competition max.	-	0.773	0.773	-	-	-

Table 1: Accuracy of our method in 5 fold cross validation, and comparison to other methods. Note that the sets are not exactly the same and not directly comparable.

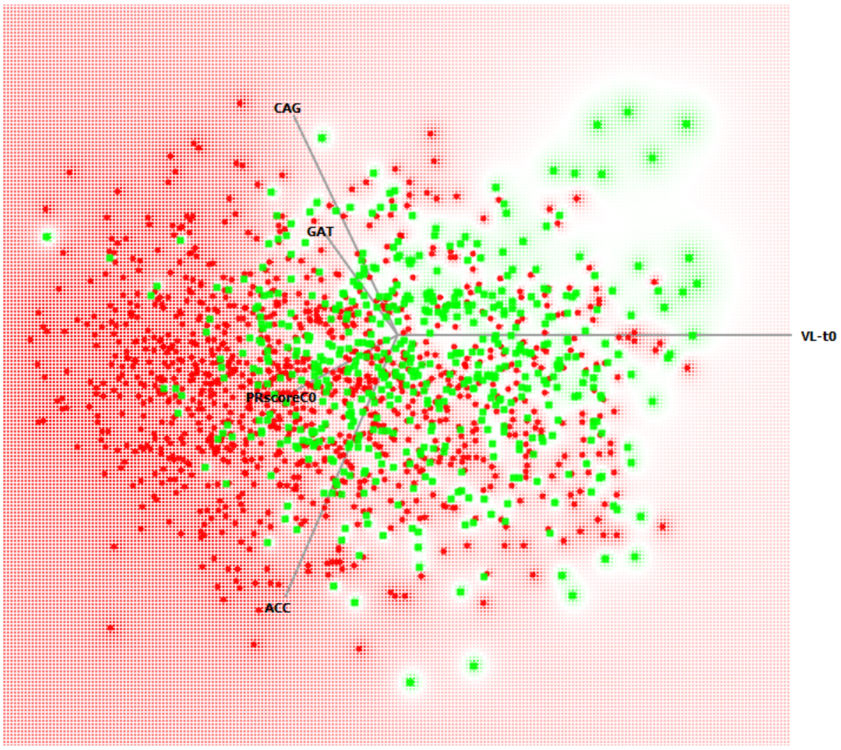

Figure 1: 2D Linear projection of 5 attributes selected by the VizRank algorithm. Red circles represent class 0 (no/little response o treatment) and green boxes represent class 1 (favorable response). The background color is the probability of a point belonging to either class, calculated by weighted k-NN algorithm. Shown axes: VL-t0 is the viral load before beginning of the treatment, PRscoreC0 is the HMM score when a sample is matched against the PR sequences profile of class 0, and CAG,GAT and ACC are the relative frequencies of those 3-mers in a sample.

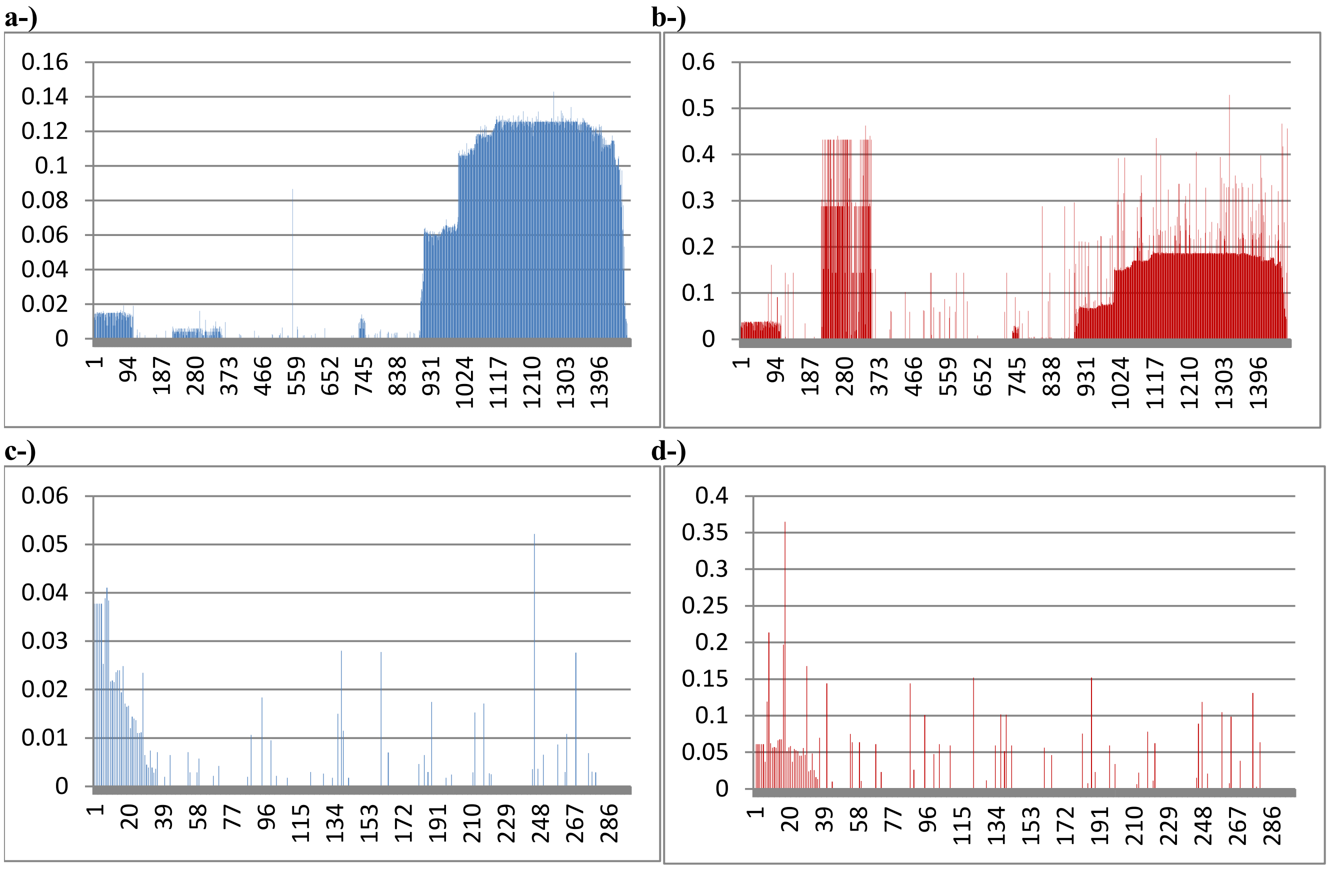

Figure 2: Information of the RT and PR sequences with respect to alignment position. a- Information gain for RT sequences, b- Gain Ratio for RT sequences, c-Information gain for PR sequences, d- Gain ratio for PR sequences.

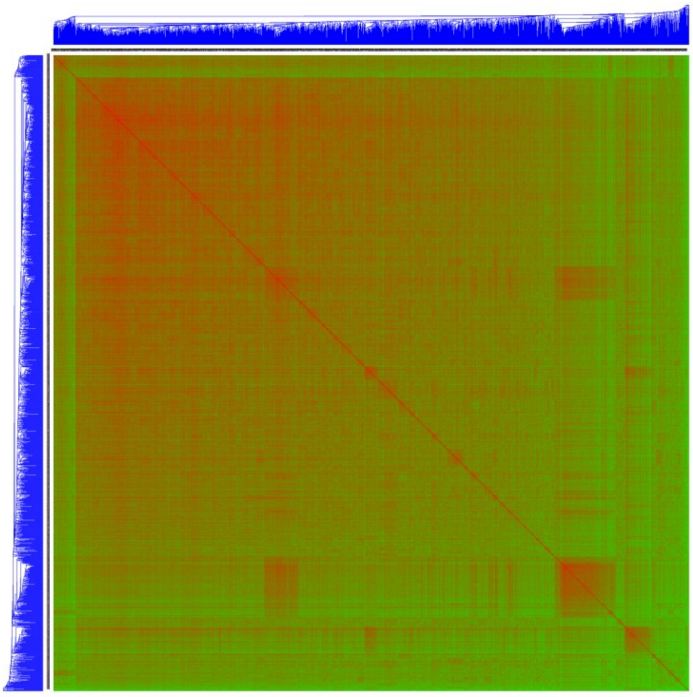

Figure 3: Heat map of the distance matrix of samples, created by taking the Euclidian distance between the feature vectors. Green indicates higher similarity. The clusters in the samples can be seen along the diagonal.