An Analysis on the Prediction of HIV Progression

Derek Chou
Computer Science

Ritchie Iu
Computer Science

Ranjay Krishna
Computer Engineering

Alan Liang
Computer Science

December 13, 2012

Abstract

This project analyses the progression of the human immunodeficiency virus by finding markers in the HIV sequence that predict a change in the severity of the infection. This severity is measured using two parameters: the viral load in a patient and the CD4 count. Using the nucleotide sequences of patients' Reverse Transcriptase (RT) and their Protease (PR) at the beginning of therapy, we test multiple machine learning algorithms using n-fold cross validation tests on the training data. The machine learning algorithms used are Multinomial Bayes Classifier, Less Naive Generative Bayes Classifier, Decision Trees, Linear Classifiers with Support Vector Machines and k-Nearest Neighbours with Needleman Wunsch dynamic programming algorithm. We discovered that there was a 62% direct correlation between the severity of the virus load and the probability of the patient improving. Through our experiments, we have ruled out certain features including codon transitions, nucleotide transitions and codon placements in patients' DNA sequences as predictive influences. However, we have also concluded that the viral load, the CD4 count and certain positions in the reverse transcriptase sequences are directly related to the progression of the HIV virus.

1 Introduction to the Problem

Human immunodeficiency virus (HIV) is a virus which leads to acquired immunodeficiency syndrome (AIDS), causing the failure of the immune system which allows other infections and cancers to thrive. It is considered a pandemic that is actively spreading across the world, and according to the Joint United Nations Programme on HIV/AIDS, as of 2009 nearly 30 million people have died of HIV-related causes (1).

Currently, the most effective treatment is a collection of antiretroviral drugs which simply suppress the virus instead of curing it (2). However, it is likely that the virus will evolve and bypass these treatments in the future. As a result, if we can pinpoint the gene sequences that determine whether the condition of the patient deteriorates or improves, it may lead to a better understanding of which gene sequences to target when building therapies and drugs to combat the virus.

2 The Data Set

We are given a training set with 1000 patients and will test on a test set of 692 patients. All patients in both training and test sets have protease sequence, reverse transciptase sequence, their viral load and CD4 count, and patients in the training set also have a binary variable that determines whether the patient gets better or doesn't (3).

2.1 Protease DNA Sequence

Protease (PR) is an enzyme that cleaves the peptide bonds linking amino acids together in the polypeptide chain forming the protein (4). HIV uses protease to cut the string of HIV-1 amino acids into numerous functional units (5).

2.2 Reverse Transcriptase DNA Sequence

Reverse Transcriptase (RT) is an enzyme that synthesizes DNA from an RNA template (6), and is responsible for the replication of retroviruses like HIV (5).

2.3 Viral Load

The Viral Load is the indicator of the severity of the viral load, measured by the number of viral particles (RNA copies) per milliliter of blood. This variable is used to determine whether a treatment is working on a patient for a given disease. (5)

2.4 Cluster of Differentiation (CD) 4 Count

The CD4 is a glycoprotein found on the surface of T-cells, a type of white blood cell. The count of CD4 cells per milliliter of blood is used to estimate the number of white blood cells per milliliter of blood. Because HIV uses CD4 to bind itself to T-cells, a higher CD4 count paradoxically represents both a healthier patient and a higher amount of HIV reproduction. (5) (7)

3 Project Scope

This paper addresses the work conducted to measure the progression of HIV in patients by comparing patients who got better with treatment for HIV versus patients who did not. Here, getting better refers to a 10 fold reduction in the viral load in the patient's system. This paper will analyze the genetic sequences and find markers that have correlation with the severity of the disease measured by the viral load and CD4 counts. Various different machine learning algorithms will be utilized with different assumptions and draw conclusions about the virus.

4 Project Background

4.1 HIV Virus

The human immunodeficiency virus (HIV) is a retrovirus, which infects humans when it comes in contact with tissues underneath the skin. The HIV infection is generally a slowly progressive disease in which the virus is present throughout the body at all stages of the disease. As HIV grows, it gains the ability to mutate, making it resistant to previously effective treatment. It suppresses the human immune system, causing acquired immunodeficiency syndrome (AIDS). To date, the best combination of drugs for HIV have not been defined (8).

4.2 Nucleotides and Codons

Each protease and reverse transciptase sequence is made up of nucleotides. The four basic nucleotides are Adenine('A'), Cytosine('C'), Guanine('G'), Thymine('T'). However, there are eleven other nucleotide types that are degenerate, that is, they can represent two or more of the basic nucleotides. This makes the analysis of nucleotides difficult as the degenarate nucleotides need to accounted for being any of the basic types.

A codon consists of three nucleotides in the sequence. In our data set, we have 443 unique codons. Codons decide which amino acid is to be produced. Figure 1 shows how nucleotides make up a RNA sequence and how they are divided up into codons.

5 Past Work on the Virus

5.1 Antiretroviral Drugs

Previously, HIV has only been treated with antiretroviral drugs. These drugs attempt to pre-

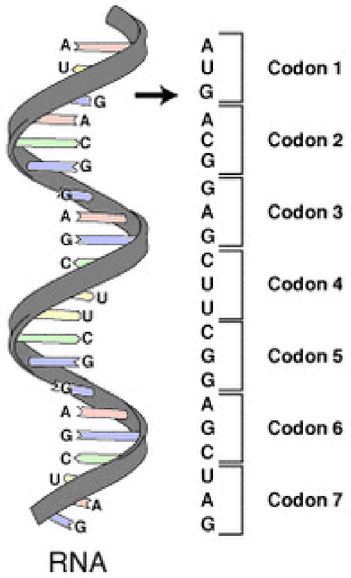


Figure 1: This image depicts the relationship between individual nucleotides that constitute a DNA sequence and thier respective codon.

vent HIV from multiplying either through blocking them from entering white blood cells, preventing them from duplicating inside, or preventing them from leaving the blood cells (9). In 1996 without the drugs, life expectancy was 10.5 years. Utilizing highly active antiretroviral therapy (HAART) has increased life expectancy to 22.5 years in 2005 (10). Although this process could suppress the viral load to low levels, it was not possible to completely cure HIV. Furthermore, HAART requires regular consumption of drugs which can be expensive and also unobtainable in some countries; it is estimated that only about half of the 34 million people infected with HIV have access to medication (11).

5.2 Berlin Patient

The development for a cure for HIV had started to slow down - the HAART treatment was effective at suppressing the virus and there was no evidence to suggest a cure was possible. However, this changed when Timothy Ray Brown, also known as the “Berlin Patient,” became the first person to be “cured” of HIV. Timothy was diagnosed with HIV in 1995, and in 2006 he also developed leukemia while living in Germany (12). Brown went through two bone marrow transplants, the second of which he received from a donor whose cells had genetic mutations that did not contain the CCR5 receptor that is necessary for HIV to replicate (13). Since then Brown has not taken an antiretroviral drugs and has been tested as HIV negative (12).

5.3 Codon 184 Mutations

Researchers have determined the importance of codon 184 in the Reverse Transcriptase DNA. Mutations for codon 184 were tested against 3TC in Simian Immunodeficiency Virus (which scientists believe HIV is derived from) which showed resistance against viral replication (14). Furthermore, in a separate study, 184 was tested in HIV which showed resistance against ddITP (15).

6 Machine Learning Approaches

The sections below describe the various different assumptions that were made about predicting the HIV progression. This document will list out details of our hypothesis and the algorithms that were used to tackle each of them. It will conclude each section with the results that were obtained from each approach.

Mulitple random assignments to patients were alloted to estimate a baseline prediction accuracy. The average accuracy was about 50%. This result was expected since out of the 692 patients in the test set, exactly half of them got better while the other half got worse.

7 Codon Transitions

7.1 Overview

Our initial hypothesis assumed that we can classify patients based on the transitions of codons in their protease or reverse transcriptase. We used a Less Naive Bayes Algoithm and calculated the probability of transitions between codons. We also calculated how often a codon appeared in patients that got better as opposed to patients that got worse. With these given probabilities, we use the following formula to classify the patient in the test set.

$$h_{less}(x) = \underset{y\in\{+1,-1\}}{\operatorname{argmax}} \left\{ P(Y=y)P(W_1=w_1|Y=y)\prod_{i=2}^{l} P(W_i=w_i|W_{i-1}=w_{i-1}, Y=y) \right\}$$

Figure 2: Less Naive Bayes algorithm decision rule

7.2 Outcome

We ran the algorithm initially considering just the protease sequence and ignoring the reverse transciptase and vice versa. First, we conducted a 10-fold cross validation on the training set and attained a 98% accuracy. However, we got a accuracy of 41% in the test set for both sequences. So, we decided to include both the sequences when calculating the probabilities. Unfortunately, this did not improve our accuracy higher than 42%. We concluded that codon transitions are not a satisfactory indicator of which patients' disease would progress further.

8 Nucleotide Transitions

8.1 Overview

Our next attempt was to go a step further and investigate the dependence between individual nucleotides instead of grouping them into codons. We used the same algorithm and concentrated on nucleotide transitions and occurences in patients instead of codons. We were hesistant towards using this approach since there are only 15 distinct nucleotides, making their occurences in large numbers in patients inevitable. Since we had more patients who got worse instead of patients that got better, our data was skewed towards favoring one classification over another. We had to balance this out.

8.2 Outcome

This algorithm did not provide a better error rate. We were still at an accuracy close to 41%. The data that we had was skewed towards predicting that patients would not get better since there were more occurences of nucleotides in that class. Even if an equal number of both the classes were considered, we had to account for patients with variable length nucleotide sequences. We learned from these two approaches that the transitions of codons or nucleotides were not a good indicator of whether a patient will improve.

9 Multinomial Bayes (Generative) Approach

9.1 Overview

Given that the training and test data have 443 unique codons, we wanted to address whether some particular codons exist only in patients that got worse versus codons that exist in patients that got better. To measure this hypothesis, we used a Multinomial Bayes Algorithm and measured the occurences of each of the codons in each class. To classify them, the following algorithm was used.

$$h_{naive}(x) = \underset{y\in\{+1,-1\}}{\operatorname{argmax}} \left\{ P(Y=y)\prod_{i=1}^{l} P(W=w_i|Y=y) \right\}$$

Figure 3: Multinomial Bayes algorithm decision rule

9.2 Outcome

Using this approach, we improved our outcome to 49%. Unfortunately, we were still below the baseline of 50%. But this allowed us to conclude again that there are no codons which are predominant in one type of patient. Codons are evenly distributed amongst all the patients.

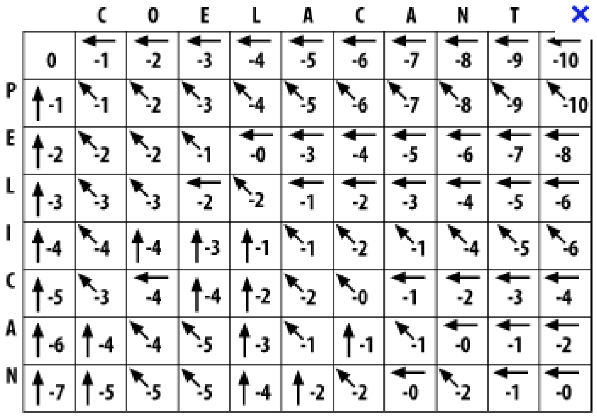
10 Needleman-Wunsch Sequence Alignment

10.1 Overview

Needleman-Wunsch algorithm is a sequence alignment algorithm used to compare two DNA sequences. It is a dynamic programming algorithm where two sequences are arranged such that they achieve the highest similarity score. It allows for gaps within the sequence at the codon (protein) level.

The protease and reverse transcriptase were converted into their corresponding codon sequences and then the codons were translated to their protein translations. Next, each patient was compared against all other patients. To compare any two sequences, a table is built where each entry, represented by the row and column index: (i,j), represents the score if the first sequences were of length i, and the second sequence was of length j.

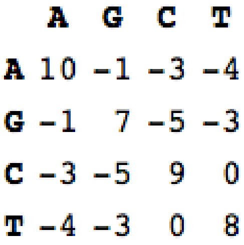
A negative score is accounted for every gap in the sequence and a positive one on pairing two proteins up. Once constructed, the table looks like the following:



		C	O	E	L	A	C	A	N	T	✕
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6	↖ -7	↖ -8	↖ -9	↖ -10
E	↑ -2	↖ -2	↖ -2	↖ -1	← -0	← -3	← -4	← -5	← -6	← -7	← -8
L	↑ -3	↖ -3	↖ -3	← -2	↖ -2	← -1	← -2	← -3	← -4	← -5	← -6
I	↑ -4	↖ -4	↑ -4	↑ -3	↑ -1	↖ -1	↖ -2	↖ -1	↖ -4	↖ -5	↖ -6
C	↑ -5	↖ -3	← -4	↑ -4	↑ -2	↖ -2	↖ -0	← -1	← -2	← -3	← -4
A	↑ -6	↑ -4	↖ -4	↖ -5	↑ -3	↖ -1	↑ -1	↖ -1	← -0	← -1	← -2
N	↑ -7	↑ -5	↖ -5	↖ -5	↑ -4	↑ -2	↖ -2	← -0	↖ -2	← -1	← -0

Figure 4: Needleman-Wunsch Algorithm Table comparing two DNA sequences

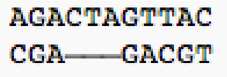
The positive scores for aligning two proteins is provided by past research through a similarity matrix like the one below:



	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

Figure 5: A similrity matrix comparing nucleotides

For example, the best alignment for "AGACTAGTTAC" and "CGAGACGT" is shown below:



```
AGACTAGTTAC
CGA---GACGT
```

Figure 6: The optimal alignment of "AGACTAGTTAC" and "CGAGACGT"

10.2 Outcome

Using the above algorithm to compare two patients, a k Nearest Neighbour classification algorithm was able to determine whether the HIV virus was going to spread. This implementation allowed us to break past the baseline and achieve a 51% accuracy. Similarity between patients, however, was not a strong measure of HIV progression.

11 Linear Classifier

11.1 Overview

Because we had not yet included viral load or CD4 count into our models, we decided to examine how these values could be used to aid our predictions. Using the primal batch perceptron algorithm, we created a linear classifer for these two variables and also included a bias variable. Since we used the batch algorithm, we iterated over the entire training sample multiple times. We wanted to see how different number of iterations affected the accuracy on the test data. In addition, we ran SVM lite with a soft margin (c equal to 0.05) on the two variables to account for the fact that classification was not solely dependent on the two variables.

11.2 Outcome

We varied the number of iterations and found that this affected how strong our predictions were. As we increased the number of iterations, we increased our accuracy. Here is a table showing the different accuracies we received for different numbers of iterations. Our best result for perceptron was at 50,000 iterations, where we got 55.3% accuracy. Afterwards there was a decline. When we ran the SVM, we received 57.1%, which was an improvement over the perceptron.

Number of Iterations	**Accuracy**(%)
10,000	50.3
50,000	55.3
100,000	55.2
500,000	53.0
1,000,000	52.9
SVM	57.1

Table 1: `Accuracy given different number of iterations of the perceptron algorithm`

To see whether the perceptron algorithm would terminate, we ran it 100,000,000 times but allowed it to stop if there were no misclassifications in one cycle. When we ran the algorithm, there was no early termination. Thus, we do not know at what point the perceptron would converge, or whether it does at all. If the algorithm does not terminate, this means the data is not linearly separable which indicates that there is a better way to formulate our prediction. However, the improved accuracy from previous methods show that viral load and CD4 values have a correlation with prediction values. In fact, with a glance at the training data, we observed that examples with higher viral loads tended to be classified as 1. In addition, examples with low CD4 values tended to be classified as 1, although the correlation with CD4 was not as strong as the one for viral load.

12 Information Gain on Codon Locations

12.1 Overview

Focusing on the viral load and CD4 count helped us realize the potential in analyzing individual data values. We decided to look at each individual codon location and see which was most useful. We hypothesized that there were only certain codon locations that were important in classifying, which is why our previous attempts in comparing the DNA as a whole were not very successful. Thus, we set out to find the most important codon locations, incorporate what we learned about viral load and CD4, and construct a decision tree to make our predictions.

12.2 Algorithm Implementation

We only looked at the RT DNA because a lot of the training examples were missing a PR DNA sequence. In order to determine which codon positions in the RT DNA were most useful, we calculated information gain for each of them and ranked them from greatest to smallest. We built our decision trees by splitting on the codon positions that gave the highest information gain. The highest index positions seemed to give the highest information gain, but upon further inspection this is because only a couple of the training examples reached that length and all of them happened to be one class. We decided it was not useful to only cover a couple of instances out of a thousand, and removed those highest position indices from our information gain ranking. After that, for the positions with highest information gain, we looked at the possible codon values and the percentage of examples with those codon values that resulted in negative and positive classifications. We used 20% as our base for percent of positive examples because that is roughly how many examples in the training data are classified as positive. We concluded that codons with a higher percentage of positive examples could be used to predict that an instance was positive. This is how we selected which codon values to split on for a given position number.

12.3 Outcome

We constructed several diferent decision trees. We created an initial decision tree using only RT215, which was the position with highest information gain. This notation means we used the reverse transcriptase DNA at position 215. In this decision tree, if the example's value at position 215 was TTT, TTA or TTC, and the viral load was greater than or equal to 4.5, it was classified as 1, and otherwise 0. We knew from before that higher viral load meant higher chance of recovery. This classification gave us an accuracy of 57.8%.

Next, we decided to experiment with RT184. RT184 was the position with the second highest information gain and we chose it because it had a lot fewer codon possibilities than RT215. We noticed that different codon values had different distributions of positive and negative results. To improve our classification, for codons with a higher percentage of positive examples, we gave a lower threshold for viral load because it was more likely to indicate a positive example. Thus, for our tree, if the example's value at position 184 was YGT, TAT or CAT, and its viral load was greater than or equal to 4.4 OR the value at 184 was CGT and its viral load was greater than or equal to 5, it was classified as 1. With this tree, we got 60.4% accuracy.

For our next tree, we wanted to include more levels. For the first level, we made the same division as we did for the RT184 tree. The examples that fit the criteria were classified as positive. For the other examples, we gave an additional check. If the example's RT215 was equal to CAC, TTT, or TTA and its viral load was greater than or equal to 4.3, it was classified as positive. This tree gave us 61.1%. As a result, we found that by deepening the decision tree, we were able to make better classifications. Finally, we expanded our decision tree further by including RT98. We also lowered our viral load threshold values so that more items could be classified as positive. Here is the structure of our tree.

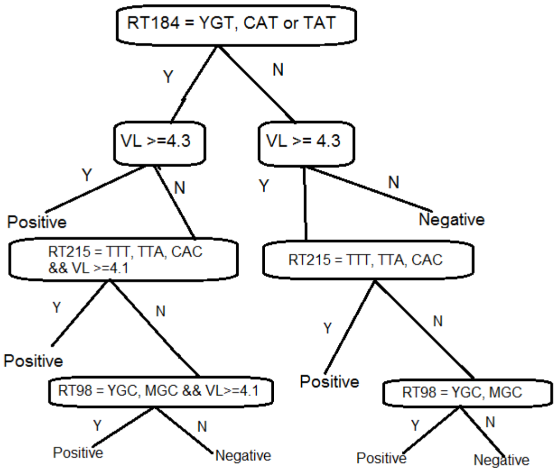


Figure 7: Our Decision Tree

In addition, we added two separate trees. One classified items with RT215 = TAC and viral load greater than or equal to 5 as positive, and the other classified examples with RT184 = CGT and viral load greater than or equal to 5 as positive. RT215 = TAC and RT184 = CGT did not have high percentages of positive examples, but we did not want to completely classify all of them as negative which is why we added them separately with a high viral load threshold. With this group of decision trees, we got an accuracy of 62.1%. Overall, accuracies from decision trees were our highest results, and confirmed our belief that certain codon positions were useful in predicting patient outcomes.

13 Future Work

One shortcoming that we encountered was the limited amount of data. The training data was skewed so that only around one-fifth of the training examples had positive classifications. In addition, we could not determine whether our predictions were skewed towards false negatives or false positives since our test data was limited by having to test against databases online. In the future, we would like to gather more data to strengthen our predictions.

Finally, we would like to do more experimentation with decision trees. We can continue building the decision trees to the point of overfitting and prune the lowest nodes to reach an optimal predictor.

14 Results

Methods	**Accuracy**(%)
Codon Transitions	41.00
Nucleotide Transitions	41.00
Multinomial Bayes Approach	49.00
Needleman-Wunsch Alignment	51.00
Linear Classifier	55.30
Support Vector Machines	57.10
Decision Trees	62.1

Table 2: `The accuracy results of all the difference algorithms that were used.`

15 Conclusion

Through a series of hypothesis formation and experimentation, we were able to identify key elements that determined the progression of the HIV virus and eliminate features that were not relevant. Our original idea of transitions between codons being a deciding factor was proven false by the low accuracy rate on the test set. We determined through the Less Naive Bayes approach that transitions between codons and even nucleotides are not a good indicator of whether a patient will progress onto AIDS. Using the Multinomial Bayes we also concluded that out of the 443 unique codons in out data set, there were no specific codons that occured in patients that improved versus patients whose conditions deteorated.

Migrating from a Generative approach to a Discriminative one, we attempted to create a measure for comparison between two patients using sequence alignments of their respective protease and reverse transcriptase DNA sequences. Using a comparison, we would be able to predict which patients were most alike though a k-Nearest-Neighbor algorithm. For the sequence alignment, we used Needleman-Wunsch. Unfortunately, similarity between patients was not a good indicator either.

Next, we concentrated on viral load and the CD4 count linear classifier. Our intuition that the higher the viral load, the more likely a person would improve with medication and the higher the CD4 count, the lower likelihood of recovery turned out to be a valid assumption. We arrived at a good indicator with a 55.3%. By using a soft margin SVM to account for errors, we improved further to 57.1%.

Our final approach relied on specific locations in the reverse transciptase of the patients. Some specific locations of codons were good predictors. We exploited these features by calculating the information gain on each of the locations on out test set and build a decision tree for classification that managed to improve our results to 62.1%.

Using our experiments, we have ruled out certain features including codon transitions, nucleotide transitions and codon placements in patients as predictive influences. However, we have also concluded that the viral load, the CD4 count and certain positions in the reverse transcriptase sequences are directly related to the progression of the HIV virus.

16 Acknowledgements

It was exciting working a project with real data from actual patients. Our team wanted to conduct research on an unsolved problem that influences many people. We ran across the idea of working on HIV progression prediction when we found articles about the current research that is being carried out in this field. It was a challenge starting on the project as none of us have any background in the HIV retrovirus or biology in general. We started slow, by reading articles on the topic and finally planned out a series of approaches we wanted to test.

We would like to thank Professor Thorsten Joachims and his staff, without who, we would not have been able to arrive at the above conclusions.

References

[1] Joint United Nations Programme on HIV/AIDS, "UNAIDS 2010 Global Report: Fact Sheet," 2009. http://www.unaids.org/documents/20101123_FS_Global_em_en.pdf.

[2] National Institute of Allergy and Infectious Diseases, "HIV/AIDS," 1998. http://www.niaid.nih.gov/topics/hivaids/understanding/treatment/pages/default.aspx.

[3] Kaggle, "Predict HIV Progression: Data Files," April 2010. https://www.kaggle.com/c/hivprogression/data.

[4] N. L. of Medicine Medical Subject Headings, "Peptide Hydrolases," 2011. http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term=Proteases.

[5] Kaggle, "Predict HIV Progression: Background," April 2010. https://www.kaggle.com/c/hivprogression/details/Background.

[6] N. L. of Medicine Medical Subject Headings, "RNA-Directed DNA Polymerase," 2011. http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term=RNA+Transcriptase.

[7] S. Ryu, A. Truneh, R. Sweet, and W. Hendrickson, "RCSB Protein Data Bank - RCSB PDB - 1CDH Structure Summary," 2011. http://www.rcsb.org/pdb/explore/explore.do?structureId=1cdh.

[8] E. S. Daar, "Human Immunodeficiency Virus (HIV Management)," *MedicineNet (online)*, March 2012. http://www.medicinenet.com/human_immunodeficiency_virus_hiv_aids/article.htm.

[9] Health Economics and H. Research Division, "How arvs work (part 3 of 3)," November 2011. http://www.youtube.com/watch?v=UuszI8l0B2w.

[10] T. Andrews, "Five Facts About HIV — LIVESTRONG.COM," May 2010. http://www.livestrong.com/article/115336-five-hiv-aids/.

[11] E. Loury, "Scientists make curing HIV a priority," 2012. http://articles.latimes.com/2012/jul/23/science/la-sci-hiv-cure-20120724.

[12] T. Salter, "Timothy Ray Brown, 'Berlin Patient,' and His Doctor are Convinced HIV cure is Real," 2012. http://www.huffingtonpost.com/2012/09/13/timothy-ray-brown-hiv-cure-berlin-patient_n_1881004.html.

[13] G. Hütter, D. Nowak, M. Mossner, S. Ganepola, A. Müßig, K. Allers, T. Schneider, J. Hofmann, C. Kücherer, O. Blau, I. W. Blau, W. K. Hofmann, and E. Thiel, "Long-Term Control of HIV by CCR5 Delta32/Delta32 Stem-Cell Transplantation," *The New England Journal of Medicine*, 2009. http://www.nejm.org/doi/full/10.1056/NEJMoa0802905.

[14] E. Cherry, M. Slater, H. Salomon, E. Rud, and M. A. Wainberg, "Mutations at codon 184 in simian immunodeciency virus reverse transcriptase confer resistance to the (2) enantiomer of 29,39-dideoxy-39-thiacytidine," *Antimicrobial Agents and Chemotherapy*, 1997.

[15] P. Boyer and S. Hughes, "Analysis of mutations at position 184 in reverse transcriptase of human immunodeficiency virus type 1.," *Antimicrobial Agents and Chemotherapy*, 1995.