A Mathematical Model for Assessing Genetic Damage on HIV Populations after Anti-Retroviral Therapy

(BU-1506-M)

by

Ileana Borjas
Universidad National Autónoma de México

Meera Lea Pradhan
University of Texas

Magnon Ivan Reyes
Florida International University

Kenneth Jeremy Spencer
Texas A&M University

August 1998

Abstract

AZT, an anti-retroviral drug, kills a large proportion of HIV in a patient's body. Those killed tend to be "highly fit"; that is, they are well adapted to the envinonment of the body and those that survive are poorly fit. As time passes by, the small proportion of strains that survive the medication have a chance of mutating into strains of higher fitness. From this phenomenon, we find a unique angle to analyze these dynamics. Instead of the perspective of the population of HIV strains. Combining genetic algorithms and difference equations, we attempt to assess the genetic damage of one drug on the future generations of survivors. We use the model of difference equations to compare the viral load of the current generation to its predecessors. The genetic algorithms allow us to analyze strains of DNA in terms of binary sequences instead of nucleotides. In the simulations we can analyze the long term behavior of the population against a drug. The goal is to describe a therapy that prevents the population of HIV from exploding.

Introduction

HIV (Human Immunodeficiency Virus) is one of the most extensively researched retrovirus in the scientific community. However, finding a drug or a combination of drugs (a.k.a. cocktails) to kill this entity has been hitherto unsuccessful (Mann et al). The main reason that HIV is difficult to eradicate is because it adapts itself to the environment of the body by rapidly mutating into strains of 'higher fitness'. Their ability to survive in the environment is a function of their DNA composition; those that have certain mutations are able to adapt to a changing environment, and are considered more fit. Strains of higher fitness can replicate quickly and survive best in the body. When faced with resistance, such as a drug, HIV can 'outwit' the opposition. For instance, by changing the sequence of a receptor gene, which is a gene that makes the virus more virulent, it increases its resistance to drugs. Researchers have found that a population of HIV under continuous treatment of a drug like Azidothyamine (AZT, nucleotide analogue) become resistant over time (Kirschner). In this way, long term treatments are very problematic to implement. Therefore, the goal of the research community is to find a medicine that hits the population of HIV soon after infection with such force that it is unable to recover. With this, HIV will not have time to develop resistance and survive, and long-term treatment will be unnecessary (Bartlett et al).

To begin, we examine the effects of a drug after it has been introduced into a sample of HIV. There will be a small proportion of the strains that will survive (Djurisic). A good medication targets strains of high fitness. If the medication kills the target strains, the survivors should tend to have a low average fitness inside the body. (In the environment, higher fitness corresponds to higher frequency and the opposite holds true as well). Reproduction will be difficult for these survivors. HIV mutates rapidly, though, and subsequent generations will have a higher fitness than the preceding ones. After a few generations some strains can evolve to a fitness in which the population will grow explosively and become a threat again. Given the relatively high reproduction and mutation rates of the virus, population and fitness may increase rapidly; in this way, simply looking at the volume eliminated by a drug ignores a large part of the picture. We redefine the idea of damage to a population of individuals in terms of genetic damage. This can be measured by including a gauge of the average fitness of the surviving population. With an appropriate treatment, the population's overall fitness and volume would be so low that it would not recover. Analyzing the dynamics of the viral population after a drug has been introduced is critical because it will help to decide how strong a drug dosage should be and what timing is required in order to prevent the population from recovering and exploding.

We decided to address this issue with the help of mathematical models and genetic algorithms (GA's). Our objective is to come up with different and innovative methods to analyze data after drug administration from the perspective of viral DNA sequences rather than a population of cells. Instead of using DNA sequences of nucleotides, such as GATTACA, we use binary codes of 0's and 1's in order to implement the information we have into a computer program. We use this technique to perform an experiment in which we apply a drug to a number of sequences, or strains, of desired length. From this simulation we attempt answer the following question:

If a medication kills a certain number of strains in the sample population, how can we measure the damage done to the population? We index how successful a medication is by observing which strains die and in what proportions, and then we implement both deterministic analysis and stochastic processes to determine the future of the surviving virus.

This kind of approach will be useful in determining a timeline and potency guidelines for the application of medicine such that HIV will not have the opportunity to mutate out of reach of a medication. In real experiments, researchers are interested in knowing how long they have to wait before applying a second drug. The reason this is important is because HIV tends to mutate to a highly fit entity, and once this has happened treatment will fail. By doing this analysis we will get a better idea when to apply the next dosage and to eliminate the virus once and for all.

It has been seen that AZT patients who receive another drug, ddI, after two years of being treated with AZT, become resistant to both; however, they now are more resistant to ddI and less resistant to AZT. Researchers have been observing different control groups to address this issue and they have found that if they stop applying drugs for two months, HIV does not reduce its fitness in response to the AZT or to the ddI.

Methodology

We examine the concept of genetic damage with both deterministic and stochastic tools. The tools we develop could be applied to any virus with behavior similar to HIV. Once we develop these tools, they can be applied to a population of viruses. Data can be inputted, and both models will demonstrate the behavior of the virus after a single treatment of a given medication.

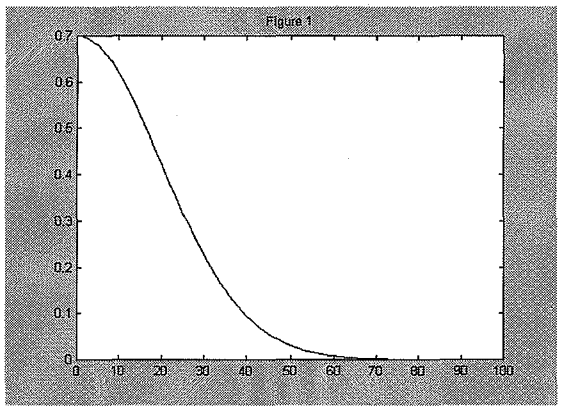
The Stochastic Approach

We begin the stochastic approach by considering a population of strains of a generic virus. This population is taken to be a sample population, representative of the real population of viruses. We assume that we can sequence each of these strains. Instead of representing the sequences of DNA as strings of nucleotides, we think of it as strings of 0's and 1's. It should be noted that one could keep the string's original information, developing a simple two digit representation for each base pair, to accurately represent the sequence (i.e. 00 for A, 01 for C, 10 for G and 11 for T). We simplify our model and consider that sequences are composed by two bases 0 and 1 exclusively.

With this sample population, we begin to construct our tools. First, we define the distance between two sequences as the number of places they have a different base, and with this definition we developed an index to measure the "spread" of the population (See Appendix B,A). The measure of spread that we developed is:

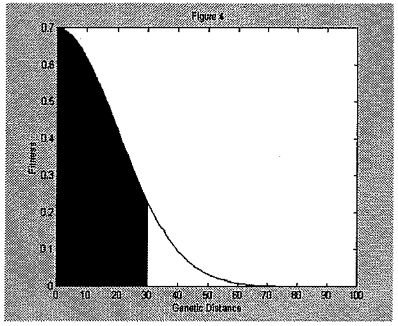
$$d^* = \frac{\sum_{i=1}^{n} d(a_i, a_*)^2}{n-1}$$

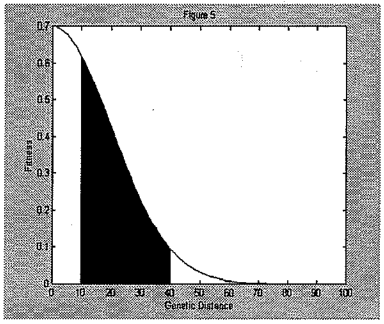
To calculate the above formula, one must identify the 'center strain' in the sample, the string whose sum of distances to the other strings is the smallest of all strings (See Appendix B,A). This is also the most frequent strain. We can then examine the population frequency versus the distance each strain is from the center. We hypothesized that the population should distribute itself from the center strain according to a binomial distribution $\binom{n}{i}p^i(1\text{-}p)^{n-i}$ with i as the number of mutations and n the length of the sequence. The following is a picture of the expected distribution, the normal, with zero as the center strain.

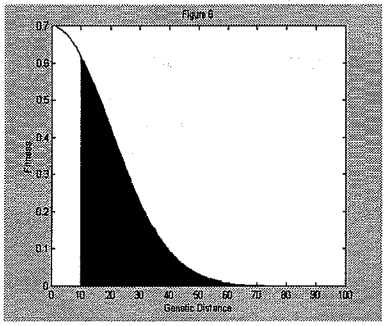


To verify, we then take our random population generated by the computer and run it through a genetic algorithm (See Appendix A). The genetic algorithm is designed to take the population of strains given a specific fitness function and a specific mutation rate, and evaluate the fitness of the strains. As we run simulations with the genetic algorithm, we can see the evolution of the population from a random scattering of strains to a distribution with the strains closest to the center occurring most frequently in the population.. The fitness function selects for the more fit strains, and consequently, they become the more frequent, thus closer to the center. The graph varies according to how many simulations are run and which fitness functions and mutation probabilities are inputted. In general, however, we find the hypothesis generally correct, as we have a decreasing function mimicking the normal distribution.

The general shape of this curve is very useful; if we normalize this curve, dividing by f(0) and renaming it g(x), what was previously a frequency curve can now be thought of as a fitness curve. G(x), the fitness, represents the *proportion* of the population that would survive at a given genetic distance from the center. Figure 2 is the frequency curve, while Figure 3 is its normalized counterpart, the fitness curve.







The drug has a specific genetic interval, represented above by the shaded region. That is, it can kill a window of strains of a certain distance to the center. For our purposes, we will only consider a drug which kills from the center strain as its minimum (See Figure 4). We can measure the endpoints of this window with an index (See Appendix B,B).

$$b = \max\{d(d_i, a_*) \text{: } d_i \text{ is a dead strain}\}$$

(For our purposes, in our index, 'a' would always be zero, the center strain.)

The length of the interval is relevant to the proportion of the population killed. Here, we lay the foundation for our concept of genetic damage.

Now our task is to measure this genetic damage to the population. The goal is to kill not only a large proportion of strains, but also to kill the most fit strains, those close to the center. Our index must consider both of these factors. To begin, we must construct a measure of the average distance of the surviving population. This is given by $\sum_{x=b}^{\infty} xf(x)$ where f(x) is the fitness function (See Appendix B,C). This is the weighted average of each distance, allowing us to consider both the proportion of a given strain and its distance from the center. Now that we have the average distance from the center of a surviving population, we can consider the drug's efficacy in terms of the volume it kills. We must factor in population size without making the index depend on the specific sample size. Hence, we use the proportion of the population left surviving. $\sum_{x=b}^{\infty} f(x)$ (See Appendix B,C). Therefore, if we take the product of these two terms, we have a measure of genetic damage. $\sum_{x=b}^{\infty} f(x) * \sum_{x=b}^{\infty} xf(x)$. This is also a measure of the subsequent generations' probability of survival.

With this information, we now allow the surviving strains to reproduce in the genetic algorithm, and we see if they can survive or if they go to extinction. The idea here is that if the fitness is low enough, or if the proportion of survivors is low enough, or some combination of both, we may be able to drive the population to extinction. In terms of drug efficacy, this means we want to use drugs with different maximums and different interval lengths. In our simulations, we experiment with varying efficacies of a drug. Some surviving populations will die off; others will survive.

With the general approach established, one could now analyze a population of HIV. Inputting relevant data, we can experiment with different medication treatments and analyze if there is a drug which can eliminate HIV with a strong single hit. However, the problem arises when we want to expand our sample and consider larger populations of strains. To run this sort of analysis could be very time consuming, and so we turn to a new kind of analysis to complement what we have started to develop.

The Deterministic Perspective

In the second phase of our project, we develop a deterministic approach which takes the surviving strains and uses them in a difference equation (See Appendix B,E). Specifically, we classify a distribution of the populations by their distance from the center strain rather than by their individual sequence. This frees us from considering each individual strain, and instead allows us to generalize for all strains 'j' distance from the center.

Let $M[t\text{-}1,j]$ = Number of individuals at generation t-1 that are at distance j

Let M[t-1] = Vector with populations at all positions 'j' at generation t-1

To consider the subsequent population, we must multiply this distribution by the average number of offspring per strain.

k = Constant number of individuals produced by a single virus per generation (offspring). However, each offspring does not necessarily survive. Survival is determined by the fitness of each subpopulation. Therefore, we must multiply the current product by the fitness at each generation.

$f(j)$ = Fitness of one element at distance j.

Let D= diagonal matrix with fitness values in diagonal

Finally, we must consider what happens as the generations reproduce. They are subject to mutate at a certain probability.

p_{ji} = Probability that one element at distance j moves to a distance i.

We devise a matrix P which gives the probability p_{ji} that from position j, in one generation, a strain can mutate to position i. The sum of the product of all of these elements gives us the 'incoming' strains to every position.

$$\sum_{j=0}^{n} k\mathrm{M}[t\text{-}1,j]\mathrm{f}(j)\mathrm{P}ji$$

Now, we must also consider the strains which mutate out of position i. We have the same initial distribution, k and D, but the probability is changed. Since p_{ii} = probability that one element at distance i remains at the same distance in the next generation, then $1 - p_{ii}$ = probability that one element at distance i leaves that distance to move to another one in the next generation. Written in matrix form, this is [I-diag] such that 'diag' is a matrix with the p_{ii} values for its diagonal elements.

We take the product of these elements:

$$\mathrm{kM}[t\text{-}1]\mathrm{D}(\mathrm{I} - \mathrm{diagP})$$

This represents the loss of population from position i. If we subtract this product from the incoming product, we have generation 't' in terms of M[t-1]. This gives us a difference equation of matrices that simplifies such that:

$$\mathrm{M}[t] = \mathrm{kM}[t\text{-}1]\mathrm{D}(\mathrm{P} - (\mathrm{I} - \mathrm{diag}))$$

$K\mathrm{D}(\mathrm{P} - (\mathrm{I} - \mathrm{diagP}))$ is a Leslie Matrix, L[t]. We can simplify the difference equation assuming a given initial population of strains M[0] such that

$$\mathrm{M[t]}=\mathrm{M[0]}*\mathrm{L[t]}^{t}$$

To examine the long term behavior of such a system, one can examine the eigenvalues of the system to see if they are <1,=1 or >1 to see if the population goes to extinction or if it survives. If we could come up with a closed form for the values entered into P, we would be able to now generalize for the state of the system in terms of eigenvalues of the Leslie Matrix. However, since there is no closed form for this matrix, we must instead plug in values and examine the behavior of a specific system we are analyzing.

To analyze the behavior of this system that we have devised, we create a program that examines the behavior of these solutions over time. Experimenting with a P matrix specific to a given population of strains and data from HIV regarding the expected replication rate, we can plug in for different medication intervals to approximate an appropriate treatment strength for the HIV (See Results).

Genetic Algorithms Revisited

Stochastic simulations can be very useful to support analysis done on a model such as the deterministic one proposed above. We return to the idea of genetic algorithms at the end of our work to see if we can develop a tool that corroborates our determinist equation. To begin, we take a population with a defined fitness distribution, f(x), which is simply half of the normal distribution. (See Figure 1).

Notice that we eliminate the process of evolving the populations through a given fitness function by assuming the hypothetical distribution. This gives us the necessary information to run our GA. We take the initial distribution, M[t-1,j], and the fitness, f(j) as the probability of survival. We can then describe the probability that the population at position j survives as a binomial distribution, with bin(M[t-1,j],f(j)). Let this distribution be 'y_j'. With this, we can now take the matrix P from our deterministic equation, and find the probability that a given surviving strain at position 'j' (classified by its distance from the center and not its specific sequence) mutates to position 'i'. This is given by the binomial distribution, bin (y_j, p_{ji}). Therefore, the new population at position 'i' can be thought of as $\sum_{j=0}^{n}$bin (y_j, p_{ji}), with n the maximum distance from the center, also the length of the strain.

This formula does essentially the same thing as our difference equation, but we now have it in stochastic terms. If a drug kills off a certain interval of strains, then the new population is described by the above summation. As it is allowed to reproduce in subsequent generations, we observe that the fitness always increases. However, whether the population survives depends on the cumulative genetic damage done to the population. Our simulations allow us to experiment with a variety of parameters to see how we can drive the population of viruses to extinction.

Assumptions

The model has some limitations; to begin, we assumed mutations to be independent of one another. In real life, this is not completley accurate. Also, our model assumed that drug with a certain efficacy D* killed off 100% of the strains at the given genetic distances. However, in reality, a drug might only kill off a proportion of those strains. An adjustment might include some sort of 'density dependent' factor which compensate for this.

Results

With limited information, we are able to analyze our deterministic model to a certain degree of reality. It is possible to analyze a patient's blood to approximate a real fitness function specific to an HIV population. In the absence of such data, we propose an artificial fitness function. Furthermore, rather than generating a random sequence of zeros and ones for an initial population, one would need to sequence a specific gene of HIV. However, we have devised the methodology to apply easily to such data if available. We will discuss briefly the results we find with our artificial fitness function to demonstrate how our model can be analyzed.

The following data corresponds to three matrices we used for the deterministic model. P100 and P100b are matrices for a population of 100 strains, each 100 base pair long. P200 is a matrix for a population of 200 strains, each 200 base pair long. P100 has a mutation rate of .01 while P100b has a mutation rate of .001, and P200 has a rate of .01. We consider a population of strains described by an artificial 'fitness function' which is the normal distribution with a maximum of 0.7 since even the most fit strains in a given population have a probability of dying (program). We designed a program so that we can vary the standard deviation, σ, to mimic a variety of possible fitness distributions (program). We also vary the replication rate, k, up to a maximum of 1000, and apply medication. On the average, the replication rate for HIV is around 500; we used 1000 to show an extreme value. As in our simulations, the medication kills on a continuous interval from the center strain. Therefore, D* is the minimum genetic distance from the center strain that the drug has to kill in order to eradicate the virus.

P100 (p=.01)

σ	k	D*
17	250	59
17	500	63
17	1000	66
20	**250**	**70**
20	500	74
20	**1000**	**78**
25	250	88
25	500	93
25	1000	98

P100b (p=.001)

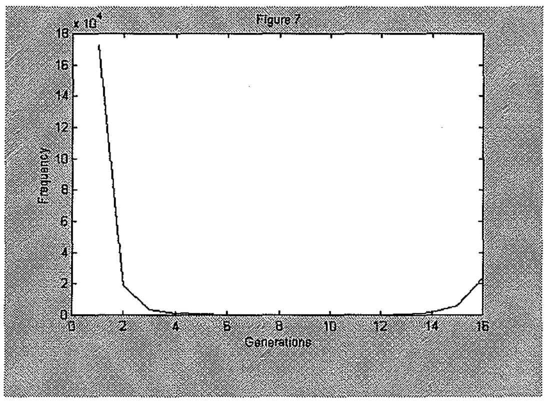
σ	k	D*
17	250	58
17	500	62
17	1000	65
20	250	68
20	500	72
20	1000	76
25	250	85
25	500	90
25	1000	94

P200 (p=.01)

σ	k	D*
35	250	116
35	500	123
35	1000	130
40	250	133
40	500	141
40	1000	149
50	250	165
50	500	176
50	1000	185

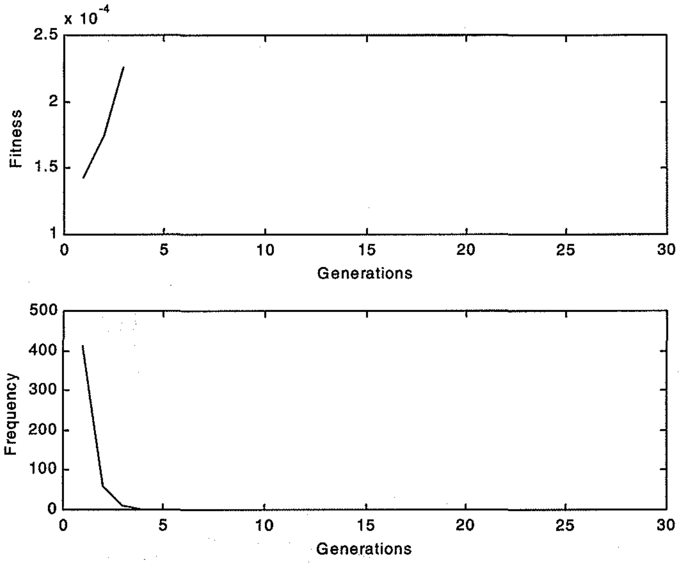
From this information, we can draw a variety of conclusions. In the P100 table above, notice the two highlighted parameters. Notice that the replication rate in the second set of values is four times greater than the other. However, D* only has to be increased by a genetic distance of 8. Since this improvement in medicine is towards the end of the maximum distance of 100, where the strains in this 8 unit region are less fit and of less frequency, it seems probable that the medication efficacy D* could be increased to eradicate the virus. Now, note that the probability of mutation of the base pairs in P100 is ten times greater than that in P100b. Remarkably, the drug efficacy necessary to kill the virus in the P100 case is not much greater than that in P100b. For instance, the parameters in the two tables for σ=20 show that for each k, D* only has to be improved by 2 units.

The following figure was obtained from the P100 matrix, and parameters: k=500, σ=20, and D*=73. When considering the D* values in the above tables, one must also consider the population distribution as genetic distance increases. This is the idea of the genetic damage index developed earlier. In this example, the population at generation 0 is 6.1098e+008, and after one generation, the population decreases to about 17e+4. Therefore, D* of only 73 killed 99.98% of the initial population after the first generation, and yet it survives.

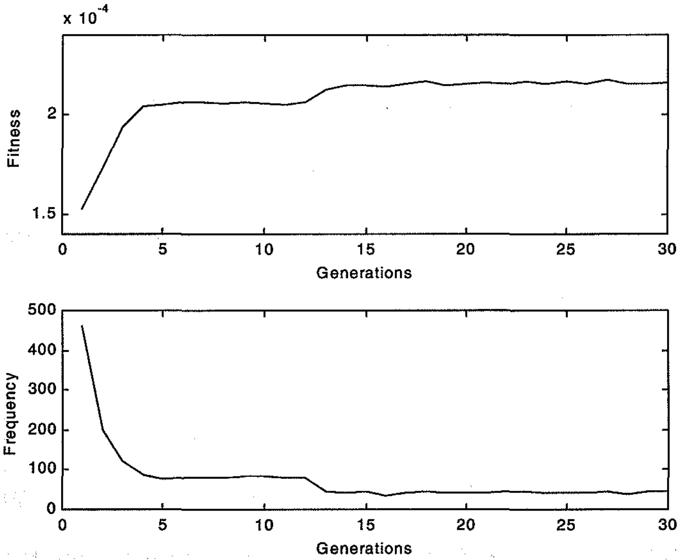


Also, a medication with efficacy better than D* will cause the virus to go to extinction. Usually, when the efficacy is below D*, the population of virus will explode. However, there are cases in which the virus population reaches a steady state. One example of this steady state that was observed is when (P100b, σ=17, k=500, D*=61).
The population rested at 4 after a few generations.

The stochastic model that we devised to correspond with our deterministic model gives a new perspective on this kind of analysis. Inputting in for various D* values, we are able to observe the two basic cases; both of the following figures demonstrate an increasing fitness, but the population survives in one and dies in the other. This makes sense, as the fitness always increases as populations evolve and adapt. The first, in figure 8, demonstrates the fact that although the fitness increases, the population can still die off if their numbers are few enough. However, in figure 9, we see a similar situation to that described previously, with a population that eventually recovers and then begins to grow.



(Figure 8)



(Figure 9)

Conclusions

The results examined above represent a few examples of how these tools can be used with real populations of HIV. The bulk of our efforts have been directed towards developing and refining these tools. In conjunction with one another, genetic algorithms and the deterministic model can be very valuable for analysis of the effects of different medications on different populations. They are both built around the fundamental concept of this project, the idea of genetic damage. This idea represents a novel approach for analyzing HIV, unlike normal cell population analysis. This will hopefully offer a unique set of tools to researchers searching for a cure to this deadly virus.

APPENDIX A
GENETIC ALGORITHMS REVIEW*

Many computational problems require a computer program to be adaptive- to continue to perform well in a changing environment.

Biological evolution is an appealing source of inspiration for addressing these problems. Evolution is, in effect, a method of searching among a huge number of possibilities for the best "solutions".

In Biology, the the diversity comes from the set of possible genetic sequences, and the desired "solutions" are highly fit organisms- organisms well adapted to survive and reproduce in their environments. The fitness criteria continually change as creatures evolve, so evolution is searching a constantly changing set of possibilities.

Furthermore, evolution is a massively parallel search method: rather than working on one species at a time, evolution tests and changes millions of species in parallel.

Viewed from high level, the "rules" of evolution are remarkably simple: species evolve by means of random variation (via mutation, recombination and other operators), followed by natural selection in which the fittest tend to survive and reproduce, thus propagating their genetic material to future generations.

Genetic algorithms (GAs) are search algorithms based on the mechanics of natural selection and natural genetics. Genetic algorithms were invented by John Holland in the 1960's and his goal was to design algorithms to formally study the phenomenon of adaptation as it occurs in nature and to develop ways in which the mechanisms of natural adaptation might be imported into computer systems.

GAs is a method for moving from one population of "chromosomes" (e.g. strings of 1's and 0's, or "bits") to a new population by using a kind of "natural selection" together with the genetically-inspired operators of crossover, mutation and inversion.

Each chromosome consists of "genes" (e.g bits); each gene being an instance of a particular "allele" (e.g 0 or 1). The selection operator chooses those chromosome in the population that will be allowed to reproduce and on average the fittest chromosomes produce more offspring than the less fit ones.

* Mitchell, Melanie. 1996

For the purpose of our project, the only operators that are allowed are mutations. Mutations are random changes in the allele value of some locations in the chromosome.

The most important parts of a genetic algorithm are:

i) populations of chromosomes,
ii) selection according to fitness,
iii) mutations to produce new offspring.

GAs most often require a fitness function that assigns a score (fitness) to each chromosome in the current population. The fitness of a chromosome solves the problem at hand.

SUMMARY
GENETIC ALGORITHMS vs NATURAL ADAPTIVE SYSTEMS

GENETIC ALGORITHMS	**NATURAL SYSTEMS**
0, 1	**alleles**
bit (e.g (000110110))	**gene**
mutation: change in one entry of the strain	**mutation: change in one base of a sequence**
chain of bits	**chromosome**

Lets consider the next code:

A=00
T= 01
C=10
G=11

Next, if we have the sequence (ATCCGTATG) that represents a gene, then in terms of GA, we can convert that sequence into (000110101101000111) using the code above, and we would consider that a bit.

If that gene belongs to a larger sequence, lets say (ATC**ATCCGTATG**GCT), then this sequence represents a chromosome and its corresponding strain, (000110**00001101011010001111**111001), a chain of bits, is also known as chromosome.

In this case {A,T,C,G} correspond to the biological alleles and {0,1} are the possible alles in GAs.

If we have the sequence (ATC) and it mutates into the sequence (TTC), then the corresponding strain of (ATC) also mutes: in fact (000110) mutates into (010110).

APPENDIX B
DERIVATION OF THE MATHEMATICAL INDEXES AND MATRICES.

A. DERIVATION OF SPREAD.

By spread we are referrring to the largest genetic distance from the 'center strain'. The center strain is the strain whose average distance to all other strains is comparatevely smallest.

Genetic distance is defined as the number of places in which two strains are differents. Mathematically this is writen as:

$$\text{let x} = [x_1,\ x_2, ..., x_n] \text{ and}$$
$$\text{y} = [y_1, y_2, ..., y_n].$$

$$\text{then, d(x,y)} = \sum_{i=1}^{n} |xi - yi|.$$

First of all, we have to find the center strain.

To do that, we have to construct the next matrix:

Suppose we have M different strains (v,w,x,y,z), construct the matrix of distances as follows:

$$\text{d} = \begin{pmatrix} \text{d(v,v)} & \text{d(v,w)} & \text{d(v,x)} & \text{d(v,y)} & \text{d(v,z)} \\ \text{d(w,v)} & \text{d(w,w)} & \text{d(w,x)} & \text{d(w,y)} & \text{d(w,z)} \\ \text{d(x,v)} & \text{d(x,w)} & \text{d(x,x)} & \text{d(x,y)} & \text{d(x,z)} \\ \text{d(y,v)} & \text{d(y,w)} & \text{d(y,x)} & \text{d(y,y)} & \text{d(y,z)} \\ \text{d(z,v)} & \text{d(z,w)} & \text{d(z,x)} & \text{d(z,y)} & \text{d(z,z)} \end{pmatrix}$$

Since d(a,b) = d(b,a) for each a,b = v,w,x,y,z, then, d is a symetric matrix with 0's in the diagonal.

Next we are going to take the sum af all the entries in one row, for all the rows, and define d_1 as

$$d_1 = \begin{pmatrix} \sum_{i=v}^{z} d(v,i) \\ \\ \cdots \\ \\ \sum_{i=v}^{z} d(z,i) \end{pmatrix}$$

For example:
Let x=[1 1 1 0]
y=[0 1 1 0]
z=[0 0 0 1]

$$d=\begin{pmatrix} 0 & 1 & 4 \\ 1 & 0 & 3 \\ 4 & 3 & 0 \end{pmatrix}; \quad d_1=\begin{pmatrix} 5 \\ 4 \\ 7 \end{pmatrix}$$

then, we will choose as a center the sequence which has the smallest value for the sum previous defined.

In our example, the center will be y.

Now, once we have selected a center strain, we are going to calculate what spread are the strains:

In general terms, if we have n strains $(a_1, a_2,...,a_n)$ the spread index is:

$$d^* = \frac{\sum_{i=1}^{n} d(a_i,a_*)^2}{n-1}$$

where a_* is the center strain.

In this case, let a_1=x
$a_*=a_2$=y
a_3=z

$$d^* = \frac{\sum_{i=1}^{3} d(a_i,a_*)^2}{2} = \frac{1^2+0^2+3^2}{2} = \frac{10}{2} = 5$$

then, $d^* = 5$ is the spread index.

Note that when we develop the sum given in d*, there is one term that is 0. This always going to happen since a_* is itself one of the a_i's, let say a_j, then $d(a_j,a_*)^2 = 0$. Given that, we can consider that we are summing only n-1 terms so we are going to divide, take the average, by only n-1 (in our example, by 2).

B. DERIVATION OF THE ENDPOINTS INDEX.

We know that a drug has a specific genetic interval. In other words, it can kill a window of strains within certain distance from the center.

In order to know the genetic interval the drug kills, we consider all those strains that are dead. From those we take the minumum of all the $d(d_i, a_*)$, where d_i is a dead strain and a_* is the center strain. This will give us the minimum distance from which we can find dead strains. In the same way we can consider the maximum of all the $d(d_i, a_*)$ and this will give us the maximum distance where we can find dead strains. Since we are considering that the effect of the drug is continuous, this two indexes determine a well defined interval [a,b], given by:

$$a = \min\{d(d_i, a_*): d_i \text{ is a dead strain}\}$$
$$b = \max\{d(d_i,a_*): d_i \text{ is a dead strain}\}$$

This interval is the genetic interval on which the drug acts and 'a' and 'b' are its endpoints.

C. DERIVATION OF THE AVERAGE DISTANCE INDEX

We construct an index to measure the average distance of a surviving population after it has been hit with a drug. We take the fitness function, f(x), and find the weighted average of the values over the region of survivors. That is, we take:

$$\sum_{x=b}^{\infty} x f(x)$$

This allows us to weigh the proportion of individuals in the surviving population at every given distance.

D. DERIVATION OF THE GENETIC DAMAGE INDEX

A measure of genetic damage must factor in two things: the average fitness of the survivors, and the proportion of the population that is left surviving. Now that we have a measure of the average distance of the survivors, we can plug that value into the fitness function to get the average fitness. That is, we take:

$$\text{f}(\sum_{x=b}^{\infty} x f(x)) = \text{average fitness of survivors}$$

Now, to consider the proportion of the population left surviving, we consider the term:

$$\sum_{x=b}^{\infty} f(x)$$

The product of these two terms will give you a measure of the genetic damage on a population.

$$\text{f}(\sum_{x=b}^{\infty} x f(x)) * \sum_{x=b}^{\infty} f(x) = \text{Index}$$

E. Derivation of M[*t*]=M[*t*-1]L

We want to know the total population of HIV at time [*t*]: the number of individuals at generation [*t*].

In order to solve our problem, we use this approach: find the total number of individuals at generation [*t*] that are a distance i from the center (M[*t*-1, *i*]).

To solve it clearly we have to express M[*t*, *i*] in an easy form:

M[*t*, *i*] = Total number of individuals at generation t-1 that can move from distance j to distance i in the next generation, where $j = 0, 1, ..., n$. – Total number of individuals at generation t-1 that can leave distance i to move to another distance.

Writing that in mathematical terms, we have:

$i = 0,1, \ldots, n$ where n is the most one sequence can differ from another (also length of sequence).

$t = 0, \ldots, M$ where M is the maximum number of generations we are considering.

$$\mathrm{M}[t, i] = \sum_{j=0}^{n} k\mathrm{M}[t\text{-}1,j]\mathrm{f}(j)\mathrm{P}ji - k\mathrm{M}[t\text{-}1,i]\mathrm{f}(i)(1 - \mathrm{P}ii)$$

Where:

k = Constant number of individuals produced by a single virus per generation (offspring).

$\mathrm{M}[t\text{-}1,j]$ = Number of individuals at generation t-1 that are a distance j.

p_{ji} = Probability that one element at distance j moves to a distance i.

$\mathrm{f}(j)$ = Fitness of one element at distance j.

p_{ii} = Probability that one element at distance i remains at the same distance in the next generation.

$1 - \mathrm{p}_{ii}$ = Probability that one element at distance i leaves that distance to move to another one in the next generation.

$$\underset{\substack{\text{offspring of the}\\\text{elements in the}\\\text{generation } t\text{-1}\\\text{at distance } j.}}{k\mathrm{M}[t\text{-}1, j]}\overset{\substack{\text{fitness of } j \text{ times}\\\text{probability to move}\\\text{to position } i.}}{\mathrm{f}(j)\mathrm{p}_{ji}} = \text{Elements moving from } j \text{ to } i.$$

$$k\mathrm{M}[t\text{-}1,i]\mathrm{f}(i)(1 - \mathrm{p}_{ii}) = \text{Elements leaving } i.$$

Now, if we look at the sum we can figure out that it could be expressed as the product of 2 vectors: one the $\mathrm{M}[t\text{-}1]$ vector and the other one the $\mathrm{f}(j)\mathrm{p}_{ji}$ vector $j = 0,\ldots,n$.

Let $\mathrm{M}[t - 1] = [\mathrm{M}[t\text{-}1,0] \;\; \mathrm{M}[t\text{-}1, 1] \;\; \ldots. \;\; \mathrm{M}[t\text{-}1,n]]$ where $\mathrm{M}[t\text{-}1,j]$ corresponds to the previous definition.

Let

$$\mathrm{fP}' = [\, \mathrm{f}(0)\mathrm{p}_{0i} \quad \mathrm{f}(1)\mathrm{p}_{1i} \quad \cdots \quad \mathrm{f}(n)\mathrm{p}_{ni} \,]$$

We can construct with each M[t] the general matrix M:

$$\mathrm{M}' = [\, \mathrm{M}[0] \quad \mathrm{M}[1] \quad \cdots \quad \mathrm{M}[n] \,]$$

We also would want to construct a matrix B with the next form:

$$\mathrm{B} = \begin{pmatrix} \mathrm{f}(0)\mathrm{p}_{00} & \mathrm{f}(0)\mathrm{p}_{01} & \cdots & \mathrm{f}(0)\mathrm{p}_{0n} \\ \mathrm{f}(1)\mathrm{p}_{10} & \mathrm{f}(1)\mathrm{p}_{11} & \cdots & \mathrm{f}(1)\mathrm{p}_{1n} \\ \vdots & & \ddots & \vdots \\ \mathrm{f}(n)\mathrm{p}_{n0} & \mathrm{f}(n)\mathrm{p}_{n1} & \cdots & \mathrm{f}(n)\mathrm{p}_{nn} \end{pmatrix}$$

We already have a matrix P with all the probabilities of p_{ji} :

$$\mathrm{P} = \begin{pmatrix} \mathrm{p}_{00} & \mathrm{p}_{01} & \cdots & \mathrm{p}_{0n} \\ \mathrm{p}_{10} & \mathrm{p}_{11} & \cdots & \mathrm{p}_{1n} \\ \vdots & & \ddots & \vdots \\ \mathrm{p}_{n0} & \mathrm{p}_{n1} & \cdots & \mathrm{p}_{nn} \end{pmatrix}$$

and since we know all the fitness values (1 for each $j = 0, ..., n$), we would want to find a matrix D such that when we multiply DP we obtain B.

It is easy to confirm that the next matrix works as we want:

$$\mathrm{D} = \begin{pmatrix} \mathrm{f}(0) & 0 & \cdots & 0 \\ 0 & \mathrm{f}(1) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{f}(n) \end{pmatrix}$$

Now we can notice that $\sum_{j=0}^{n} k\mathrm{M}[t\text{-}1,j]\mathrm{f}(j)\mathrm{p}_{ji}$ is in fact k times the matrix product of the (t-1)th row of the the matrix M times the ith column of matrix B.

In the other hand we want to find $k\mathrm{M}[t\text{-}1,i]\mathrm{f}(i)(1-\mathrm{p}_{ii})$. ....(*)

First of all we have to notice that we only need the p_{ii} values so, let

$$\mathrm{diagP} = \begin{pmatrix} \mathrm{p}_{00} & 0 & \cdots & 0 \\ 0 & \mathrm{p}_{11} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{p}_{nn} \end{pmatrix}$$

We can construct the general matrix I-diagP where I = Identity Matrix to have:

$$\mathrm{I} - \mathrm{diagP} = \begin{pmatrix} 1-\mathrm{p}_{00} & 0 & \cdots & 0 \\ 0 & 1-\mathrm{p}_{11} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1-\mathrm{p}_{nn} \end{pmatrix}$$

and therefore $1 - \mathrm{p}_{ii}$ (from (*)) corresponds to the ith column of this matrix.

Moreover, $\mathrm{f}(i)(1\text{-}\mathrm{p}_{ii})$ could be seen as the ith column of the matrix

$$\mathrm{S} = \begin{pmatrix} \mathrm{f}(0)(1-\mathrm{p}_{00}) & 0 & \cdots & 0 \\ 0 & \mathrm{f}(1)(1-\mathrm{p}_{11}) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{f}(n)(1-\mathrm{p}_{nn}) \end{pmatrix}$$

This matrix corresponds to the matrix product of D times $\mathrm{I} - \mathrm{diagP}$.

So, to obtain $k\mathrm{M}[t\text{-}1,i]\mathrm{f}(i)(1 - \mathrm{P}ii)$ we only have to take the matrix product of the (t-1)th row of the matrix M and the ith column of the previous matrix.

Notice that such product is a scalar since the (t-1)th row of M is a 1 x (n+1) vector and the ith column of S is a (n+1) x 1 vector. When we multiply them we obtain a 1 x 1 matrix that equals a scalar.

Now, to put all that information in general terms we can calculate M[t] as:

$$\mathrm{M}[t] = K\mathrm{M}[t\text{-}1]\mathrm{DP} - K\mathrm{M}[t\text{-}1]\mathrm{D}(\mathrm{I} - \mathrm{diagP}) \qquad \ldots (**)$$

That expression results when we take all the entries in the (t-1)th row of M and we consider all the columns of D, P and (I – diagP).

Expression (**) can be factorized as follows:

$$\mathrm{M}[t] = K\mathrm{M}[t\text{-}1]\mathrm{D}(\mathrm{P} - (\mathrm{I} - \mathrm{diagP}))$$

and since

$$K = \begin{pmatrix} k & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & k & & & & 0 \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & k \end{pmatrix}$$

it commutes to obtain:

$$\mathrm{M}[t\text{-}1]K\mathrm{D}(\mathrm{P} - (\mathrm{I} - \mathrm{diagP})).$$

Remember that k is the constant number of individuals per generation. K is construct just to conserve the right product of matrices.

If we notice that K is known, P is known and therefore (I – diagP) is known, we can call KD(P – (I – diagP)) as a fixed matrix L.

In conclusion, from the previous fact, we get that

$$\mathrm{M}[t] = \mathrm{M}[t\text{-}1]\mathrm{L}$$

where

$$\mathrm{L} = K\mathrm{D}(\mathrm{P} - (\mathrm{I} - \mathrm{diagP})).$$

L is called the **Leslie Matrix**.

Note:

M[t-1] $= 1 \text{ x } (n+1)$ vector.
$K = (\text{n}+1) \text{ x } (n+1)$ matrix.
D $= (n+1) \text{ x } (n+1)$ matrix.
(P $-$ (I $-$ diagP)) $= (n+1) \text{ x } (n+1)$ matrix.

then, L $= (n+1) \text{ x } (n+1)$ matrix and therefore M[t] $= (1 \text{ x } (n+1))$ $*((n+1) \text{ x } (n+1))$ matrix which is a $1 \text{ x } (n+1)$ matrix. In conclusion, M[t] is a vector.

For references about the linear algebra related with Leslie Matrix see Balmer.

F. DERIVATION OF P.

A sequence that has a distance d from the 'center strain' has d places different than the center strain and n-d alike, where n is the length of the sequence . Let x be the random variable "number of mutations in the place the sequence is different from the center strain", and y the random variable "number of mutations in the places the sequence is the same as the center strain". Then the next generation the distance to the center strain will be

$$\text{W} = d - x + y.$$

where W $= 0, 1, 2, ..., n.$

The probability mass function of the random variable W is:

$$\begin{aligned} \text{P(W} = \text{w)} &= \text{P}(d - x + y = \text{w}) = \\ &= \text{P}(y - x = \text{w} - d) = \\ &= \sum_{i=0}^{n-d} \text{P}(y = i, x = d - \text{w} + i) \\ &= \sum_{i=0}^{n-d} \binom{d}{i} \text{p}^i (1\text{-p})^{d-i} \binom{n-d}{d-\text{w}+i} \text{p}^{d-\text{w}+i} (1 - \text{p})^{n-2d+\text{w}-i}. \end{aligned}$$

which, unfortunately, has no simpler form.

APPENDIX C
GLOSSARY.

BINOMIAL DISTRIBUTION. Probability of having i success in n choices. This probability is given by $\binom{n}{i}p^i(1\text{-}p)^{n-i}$ where 'p' is the probability of having one success.

CENTER STRAIN. In a family of strains, the strain whose average distance to all others strains is the minimum.

EFFICACY. The efficacy tells us how great the genetic damage is. The greater genetic damage of a drug on a population of strains, the better the efficacy.

FITNESS. How adapted one strain is to the body of the patient. For our purposes, the body defines fitness.

FREQUENCY. Number of strains that have the same genetic distance in a given population.

FITNESS FUNCTION. Function that express mathematically the fitness of the strains: $\frac{\text{\#of strains with same fitness value}}{\text{\#of strains with highest fitness value}}$.

GENETIC DAMAGE. Damage done to the HIV virulence capacity. The more fit the strains killed by the drug, the greater the damage.

GENETIC DISTANCE. Number of places in which one strain differs from another one.

STRAINS. Sequence of 0's and 1's, representing one element of the HIV population.

VIRULENCE. How dangerous the HIV population is. The more fit strains that are alive, the more dangerous.

Acknowledgements

The research of this manuscript has been partially supported by grants given by the National Science Foundation (NSF Grant DMS-9600027), the National Security Agency (NSA Grants MDA 904-96-1-0032 and 9449710074), Presidential Faculty Fellowship Award (NFS Grant DEB 925370) to Carlos Castillo-Chavez, Presidential Mentoring Award (NSF Grant HRD 9724850), and Sloan Foundation Grant (97-3-11). We also thank the INTEL corporation for providing the latest high performance computer and software. Substancial financial and moral support was provided by the Office of the Provost of Cornell University. We also thank Cornell's College of Agricultural & Life Sciences (CALS) and its Biometrics Unit for allowing the use of CALS's facilities. We also want to thank Carlos Castillo-Chavez and the MTBI staff for giving us the opportunity to participate in this program and for all of their kind help. We add a special thanks to Carlos M. Hernandez-Suarez for his dedication and for giving us three hours of his time each night for a whole week to teach us GAs, and especially for all his support to make this project possible.

Bibliography

Barbosa-Neto, José, et al. "Precision of genetic relationship estimates based on molecular markers." Euphytica 98 (1997): 59-67.

Bartlett, John G., et al. "Defeating Aids: What Will It Take?" Scientific American July 1998: 81-107.

Bulmer, Michael. Theoretical Evolutionary Ecology. Sunderland, Massachusetts: Sinauer Associates Publishers, 1994.

Cullen, M.R. Linear Models In Biology. New York: Ellis Horwood Limited, 1985.

Dewhurst, Steve. "HIV-1: Molecular Biology." University of Rochester. <http://www.urmc.rochester.edu/smd/mbi/grad/hiv297.html>

Djurisic, Aleksandra B. "Elite genetic algorithms with adaptive mutations for solving continuous optimization problems-application to modeling of the optical constants of solids." Optics Communications 151 (1998): 147-159.

Ed. Karn, Jonathan. HIV: A Practical Approach. New York: Oxford University Press, 1995.

Ewald, Paul W. Evolution of Infectious Disease. New York: Oxford University Press, 1994.

Goldberg, David E. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, Massachusetts: Addison-Wesley, 1989

Hastings, Alan. Population Biology: Concepts and Models. New York: Springer, 1997.

Kelly, John K. "Replication Rate and Evolution in the Human Immunodeficiency Virus." Journal of Theoretical Biology 180 (1996): 359-364.

Kirschner, Denise. "Using Mathematics to Understand HIV Immune Dynamics." Notices of the AMS Vol. 43, Number 2. Feb. 1996: 191-202.

Mansuri, Muzammil and Michael Hitchcock. "Anti-HIV therapy: A lesson in drug design." Chemtech Sept. 1992: 562-572.

Mitchell, Melanie. An Introduction to Genetic Algorithms. London: MIT Press, 1996.

Park, L. J., et al. "Application of genetic algorithms to parameter estimation of bioprocesses." Medical & Biological Engineering & Computing 35 (1997): 47-49.

Wein, Lawrence M., et al. "Mathematical Analysis of Antiretroviral Therapy Aimed at HIV-1 Eradication or Maintenance of Low Viral Loads." Journal of Theoretical Biology 192 (1998): 81-98.

Rosenberg, Richard S. "Stimulation of Genetic Populations with Biochemical Properties: I. The Model." Mathematical Biosciences 7 (1970): 223-257.

Rosenberg, Richard S. "Stimulation of Genetic Populations with Biochemical Properties: II. Selection of Crossover Possibilities." Mathematical Biosciences 8 (1970): 1-37.