Evolution and Somatic Learning in V-Region Genes

Alan S. Perelson[1], **Ron Hightower** [2] **& Stephanie Forrest**[2]

An interesting and still unanswered question about the evolution of the immune system's variable (V)-region genes is what are the selective forces that operate to maintain them. In a large multigene family with hundreds of genes that have overlapping function it is difficult to imagine how selection can operate on any particular gene, because its deletion or creation is expected to have little effect on the overall fitness of the animal. An extreme example is in the antibody V-region gene segments in which one individual expresses only a small fraction of the potential antibody repertoire, and the function of a gene segment can only be seen when that segment is joined to others to construct one of a large number of possible antibody molecules. Hood and Prahl [1] speculate that selection for multigene systems must occur at the level of the whole organism and not at the level of individual germ line genes. This has been difficult to show analytically or experimentally. Here we show by computer simulation that selection operating at the organismic level can provide the selection pressure needed to generate and maintain diversity in V-region gene families.

To study the evolution of gene libraries, we have defined a simplified model in which an individual's fitness is evaluated according to how well its expressed antibody repertoire recognizes randomly selected antigens. Our model follows the evolution of a population of 500 haploid individuals. Bit strings are used to represent both molecules and gene segments. As introduced by Farmer et al. [2], the patterns of the bits represent the shapes of molecules and determine their ability to bind other molecules. Antigens and antibodies are represented by bit strings of the same length. Molecular recognition occurs when an antibody bit string and an antigen bit string "match" each other at corresponding bit positions, with 0 assumed to match 1. The total number of matching (complementary) bits is called the match score, and is used to compute fitness.

Each individual in the simulated population has a haploid genotype comprised of four equal-size gene libraries from which N antibodies are constructed at each generation, as shown in Figure 1. These N antibodies are the individual's expressed repertoire. We assume that the individuals live in a world in which there is a large finite set of antigens, called the antigen universe. The antigen universe consists of a fixed, randomly chosen set of U strings. At each generation, each individual is exposed to K antigens randomly chosen with replacement from the antigen universe. The antigen universe thus represents a set of common pathogens with which we assume vertebrates have needed to cope over evolutionary time.

To compute the fitness of an individual we match each of the K antigens against the N antibodies in its expressed repertoire. Each antigen receives an antigen score, which is the maximum of all the match scores computed between that antigen and the expressed antibodies. The antigen score quantifies how well the immune system recognizes that particular

[1]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545 USA
[2]Dept. of Computer Science, University of New Mexico, Albuquerque, NM 87131 USA

antigen. The fitness of an individual is found by averaging the scores for the different antigens and dividing by the antigen bit string length so that fitnesses are less than or equal to 1. We have also conducted experiments in which the fitness is chosen as the minimum antigen score and obtain qualitatively similar results. Because the antigen with the lowest antigen score might not be very pathogenic, averaging antigen scores may be a preferred fitness measure. Also, each antigen to which one is exposed does not necessarily influence survival or reproductive capacity.

The effects of evolution are simulated on our binary immune system by using the genetic algorithm [3, 4]. At each generation the fitness of each individual is evaluated. A new population is then formed in which the individuals with higher fitness have more offspring than the less fit individuals, but the total size of the population is held constant, i.e., the mean number of offspring is one. Two genetic operators, crossover and point mutation, modify the genotypes in the population as the genetic algorithm progresses. Crossover combines two genotypes into a new individual, and mutation changes the bits of an individual with some small probability. The experiments reported here were conducted with a derivative of the genetic algorithm software package Genesis 1.2ucsd, [5].

Our binary immune system has a large but limited ability to recognize antigen molecules. As the size of the antigen universe (recognition task) increases and the immune system is required to recognize more antigens, its average performance degrades. Our first experiment shows that the immune system can evolve and learn about the antigen universe but that the overall recognition capability degrades as U, the size of the antigen universe, increases (Figure 2). If U is so large that the immune system cannot learn, then any antigen would be the equivalent of a random antigen. Further, if no learning occurs then any antibody created from the gene libraries would be a "random antibody." Under these circumstances, one would expect an antibody and antigen to match at 50% of their bits. Since the fitness of the immune system is evaluated by choosing the antibody in the expressed repertoire with the largest match score, the expected fitness of an immune system with an expressed repertoire of N antibodies for a random antigen is

$$Expected\ Fitness = \frac{1}{2^{NL}} \sum_{i=1}^{L} i(s(i)^N - s(i-1)^N) \ ,$$

where L is the length of the bit string, here $L = 64$, and $s(i) = \sum_{j=0}^{i} \binom{L}{j}$ is the cumulative binomial distribution. For $N = 8$ this formula predicts an expected fitness of 0.588719, approximately the asymptote seen in Figure 2. The figure thus suggests that for an immune system with a given size genome, here 512 bits, only a limited number of antigens can be learned. Here, when the antigen universe U has 32 or fewer antigens the fitness curve is substantially above the asymptote indicating that the system has learned and encoded information about this number of antigens. Interestingly, this 512-bit genome can encode information about 32 64-bit antigens, i.e. about 2048 bits.

To quantitate the evolution of the V-region libraries, we set $U = 32$ and examined the effects on the rate of evolution of varying K, the number of antigens to which each individual is exposed to per generation and N, the number of antibodies expressed per generation.

Redundancy in a gene library reduces its ability to match antigens. This suggests that an ideal library might have a maximal amount of variety in its elements. Our experiments began with all of the genes initialized to zero, so diversity needs to evolve. This choice reflects the idea that the libraries were originally created by gene duplication, which would cause the library elements to be similar until mutation caused diversification.

Figure 3 shows that the fitness of the population increases with time, measured in generations. Populations exposed to a larger number of antigens, K, evolved faster and ended with higher fitness values. The effects of varying N, the number of antibodies expressed by each individual per generation, are more dramatic. In the model, the size of the potential antibody repertoire is 4096, so the values of N in Figure 3, the expressed repertoire, represent up to 6% of the potential repertoire. The populations with higher values of N show a more rapid increase in their fitness, and obtain a higher fitness after 2000 generations. However, with 1% of the repertoire expressed the system does almost as well as with 6% expressed. As one might expect, larger expressed repertoires imply greater sampling of the genetic capabilities of the organism and lead to improved evolution. Not surprisingly, even though evolution acts on the expressed repertoire and each organism only sees a random sample of the antigens in its environment, the "potential genetic fitness" of an organism, measured by the ability of its potential repertoire to recognize all possible antigens in the antigen universe, also increases with time (not shown). The effects of sampling, i.e. that each individual only expresses a small fraction of its potential repertoire and that each individual only sees a fraction of all possible antigens, reduces the overall fitness of the population but not its ability to evolve useful gene segment libraries.

In order to recognize random antigens one would expect that gene libraries should have gene segments that are as random as possible. However, because the antigen universe contains an explicit set of antigens, the immune system should be able to exploit the features present in this set of antigen and perform better than random antibodies. To test this, we generated random gene segment libraries and asked if they would evolve to higher fitness. As shown in Figure 4, libraries with random elments also evolve and reach the same final fitness as libraries initiated by gene duplication.

Another means of diversification of antibodies is somatic mutation. Although the gene libraries are not directly affected by somatic mutation, we find that including somatic mutation can speed up the evolution of gene libraries. This appears to be a Baldwin effect [6]. An individual may have gene segments that if mutated in one or two positions give rise to high affinity antibodies for particular antigens. Such key mutations have been observed in a number of anti-hapten immune responses. For example, in C57BL/6 mice, the B cell response to the hapten (4-hydroxy-3-nitrophenyl) acetyl (NP) coupled to a protein carrier is dominated by antibodies using a heavy chain encoded by the V_H gene 186.2 linked to the D_H element DFL16.1 and any of four J_H segments [7]. A single amino acid replacement, trytophan to leucine at position 33 in the V_H gene segment leads to a ten-fold increase in affinity. This key mutation is found in about 70% of secondary response NP-binding B cells [8]. Without somatic mutation an individual carrying genes that are close to the needed ones may not have any selective advantage over individuals carrying genes that would require large numbers of mutations to improve. To study the effect of somatic mutation on the fitness of individuals carrying germline genes close to highly desirable ones, we modified our simulation so that the fitness of an individual evaluated over its lifetime has both a germline and somatic component. For each antigen the germline component is calculated as before; it is the match score of the antibody in its expressed repertoire that best matches the antigen. This antibody is then mutated in G positions and the match score recalculated. The change in score is the somatic component of the fitness. Because, somatic mutation and affinity based selection are ongoing processes within individuals during an immune response, there are multiple opportunities for antibodies to improve. Here we do not simulate the details of immune responses within an individual, only the net effect of mutation and selection. We do this by assuming that the only mutations that survive are the ones restricted to areas of mismatch between antibody and antigen. Thus if the germline encoded best antibody for a particular antigen has say 20 mismatching positions, we assume the G mutations are restricted to these 20 positions. Thus after G mutations, on average, there will be $20 - G/2$ mismatches.

We also incorporated into the model a distinction between match score and binding affinity, so that we could control the size of the effect created by mutation. Here we study an extreme example in which if the match score is below a given threshold value, the binding affinity is assigned a zero value. Above this threshold the binding affinity assumes its maximum value. Thus a single mutation can make a huge difference in affinity as is the case of the position 33 mutation in V_H 186,2. In our experiments, which used strings of length 64, the match score threshold was set to 45 bits, and affinity was used as the fitness measure.

Figure 5 shows the results of experiments for values of G between 0 and 30. Each point is the fitness of the population after one thousand generations, averaged over thirty experiments. The vertical axis shows the genetic component of the match score, averaged for the entire population (and the thirty runs). This average genetic match score is essentially the "fitness at birth," before learning takes place. The curve shows that the highest average match scores (around 43 bits) were found when G was between 8 and 12. Thus, compared with $G = 0$, there is a significant improvement in fitness when 8 to 12 mutations are allowed.

The model shows that there can be an interaction between evolution and somatic learning. This type of interaction has been called the Baldwin Effect. Around 1896, Baldwin proposed [6] that learned or acquired characteristics could become part of the genetic makeup of succeeding generations without Lamarckian inheritance. Individuals that learn or acquire useful characteristics during their lifetimes tend to survive, and, Baldwin claimed, this would cause succeeding generations to have a higher probability of acquiring the same characteristics, even though the characteristics themselves were not genetically propagated.

Our experiment illustrates a modified version of the Baldwin effect. Instead of showing that "learning guides evolution," the experiment illustrates that somatic learning can accelerate evolution. The upward slope on the left side of the curve in Figure 5 is due to the Baldwin effect. It shows that increased somatic mutation leads to an increase in evolutionary progress during the course of one thousand generations. The explanation for this acceleration of evolution is that learning rewards those individuals that are nearer to the threshold in the affinity function. Without mutation, individuals near the threshold have essentially the same poor fitness as those further away. When the binding affinity is chosen to be a linear function of match score, so that this is no longer the case, then changes in G do not improve evolutionary progress (not shown). Thus, it appears that only when somatic mutation has a large effect on fitness will evolution be affected via a Baldwin effect.

We have shown that even though antibodies are constructed randomly from a large diverse set of gene segments with each individual expressing a random subset of possible antibodies, evolution can act on the gene libraries and lead to their improvement. Selection pressure improves antigenic recognition, even though each individual only responds to those antigens it stochastically encounters and only uses that portion of its potential repertoire that it randomly expresses. Thus, although an individual may be "unlucky" in the sense that its potential repertoire may contain the information needed to recognize an antigen, but the appropriate antibody may not be expressed at the time the antigen is encountered, over time the population as a whole still improves. The evolved antibody V-region gene libraries are diverse, but not random, and they encode information about antigens commonly encountered by the population as a whole.

Using an extended model that includes somatic mutation, we also have shown the presence of the Baldwin effect. The explanation of the effect is that somatic learning allows the population to perform local search of the fitness landscape during evolution, allowing evolution to discover which individuals are nearest to the threshold of success. However, the Baldwin effect is not a universal relation between evolution and learning, and appears to be sensitive to the shape of the fitness landscape. As we discovered, the Baldwin effect disappears when the nonlinearity in the binding value function is removed.

References

[1] L. Hood and J. Prahl. *Adv. Immunol.*, **14**: 291–351, 1971.

[2] J. D. Farmer, N. H. Packard, and A. S. Perelson. *Physica* **22D**:187–204, 1986.

[3] J. H. Holland. *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, Ann Arbor, MI, 1975.

[4] S. Forrest. *Science*, **261**:872-878, 1993.

[5] J. J. Grefenstette. *GENESIS: A System for Using Genetic Search Procedures.* In *Proceedings of a Conference on Intelligent Systems and Machines.* 161–165, 1984.

[6] Baldwin, J.M. A new factor in evolution, *American Naturalist*, **30**: 441–451, 1986.

[7] D. Allen, A. Cumano, R. Dildrop, et al. *Immunol. Rev.* **96**:5–22.

[8] A. Cumano and K. Rajewsky. *EMBO J.*, **5**:2459–2468.

ACKNOWLEDGMENTS. Portions of this work were performed under the auspices of the U.S. Department of Energy. This work was supported by NIH (AI28433, RR06555), the Santa Fe Institute, NSF (IRI-9157644), ONR (N00014-95-1-0364), Interval Research Corp., and the Joseph P. and Jeanne M. Sullivan Foundation.

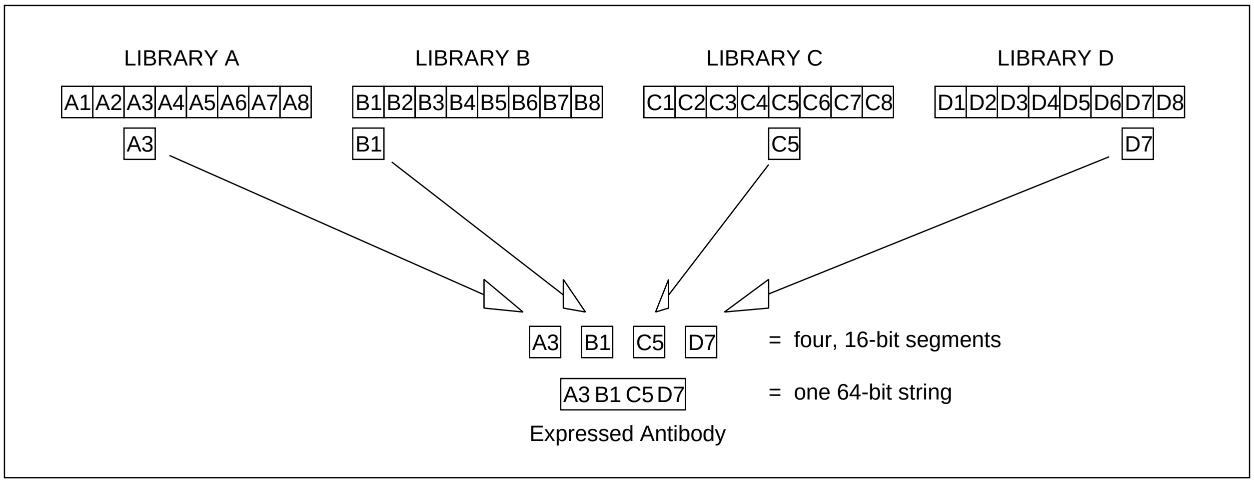

Figure 1: A simple computer model of antibody V genes and their expression. Each individual is assumed to have four V-region gene libraries, with each library containing eight gene segments, each 16-bits long. Antibodies are constructed from genes by choosing one gene segment randomly from each of the four libraries and concatenating the four 16-bit segments into one 64-bit antibody. The total number of bits in an individual genome is 4 libraries x 8 elements x 16 bits = 512 bits. The total number of different antibodies that can be made, the potential repertoire, is $8^4 = 4096$.

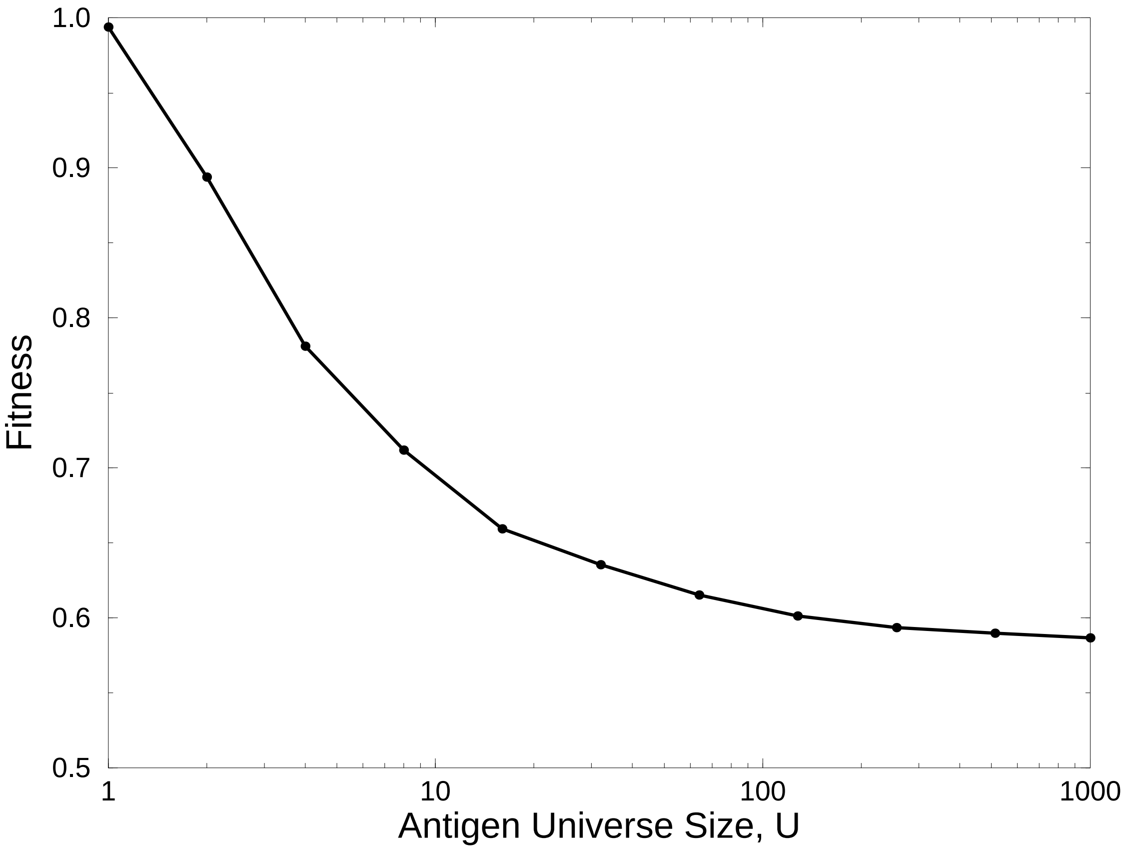

Figure 2: Evolution of the gene libraries was conducted using the genetic algorithm as described in the text. Each point in the graph is the average of the results of 30 experiments run for 1000 generations each on a population of 500 individuals. The final fitness attained is plotted versus the size of the antigen universe. In each experiment the number of expressed antibodies was fixed at $N = 8$, and the number of antigens each individual was exposed to each generation was fixed at $K = 8$. The curve asymptotically approaches a fitness near 0.589, the fitness expected if $N = 8$ antibodies are generated at random, matched against a random antigen, and the fitness of the best matching antibody recorded. Fitness values above the asymptote indicate that the gene libraries have learned and encoded information about the antigen universe.

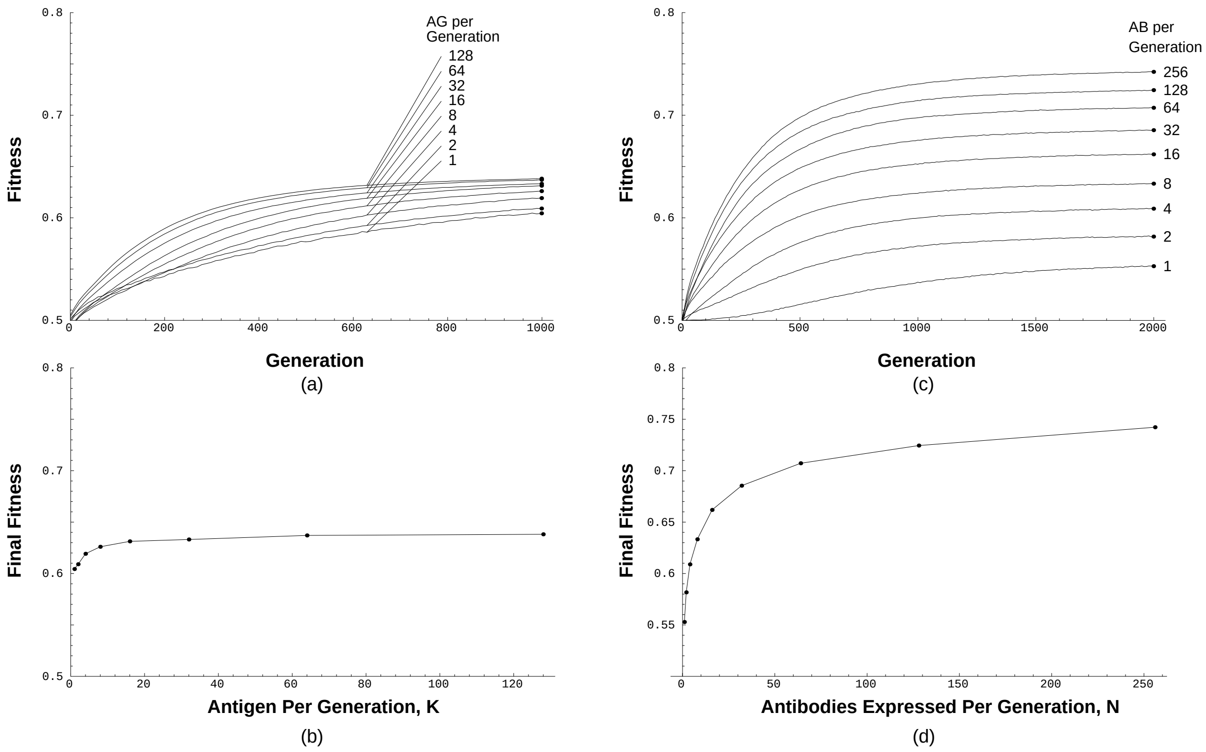

Figure 3: A population of 500 individuals was used in each experiment. The antigen universe $U = 32$. (left) Experiments comparing various rates of antigen exposure. At each generation each individual was exposed to K antigens randomly chosen with replacement from a fixed antigen universe. The expressed repertoire $N = 8$. (a) The population average fitness is shown for each experiment versus time measured in generations. All individuals started the experiment with all gene segment bits set to zero. The experiment shows that the larger K, the faster the learning. (c) Fitness attained after 1000 generations vs K. (right) As on left, except the number of antigens exposed to each individual per generation $K = 8$, and N, the number of antibodies expressed per generation was varied. The experiment shows that the larger the expressed repertoire the faster the learning and the higher the final fitness of the evolved population.

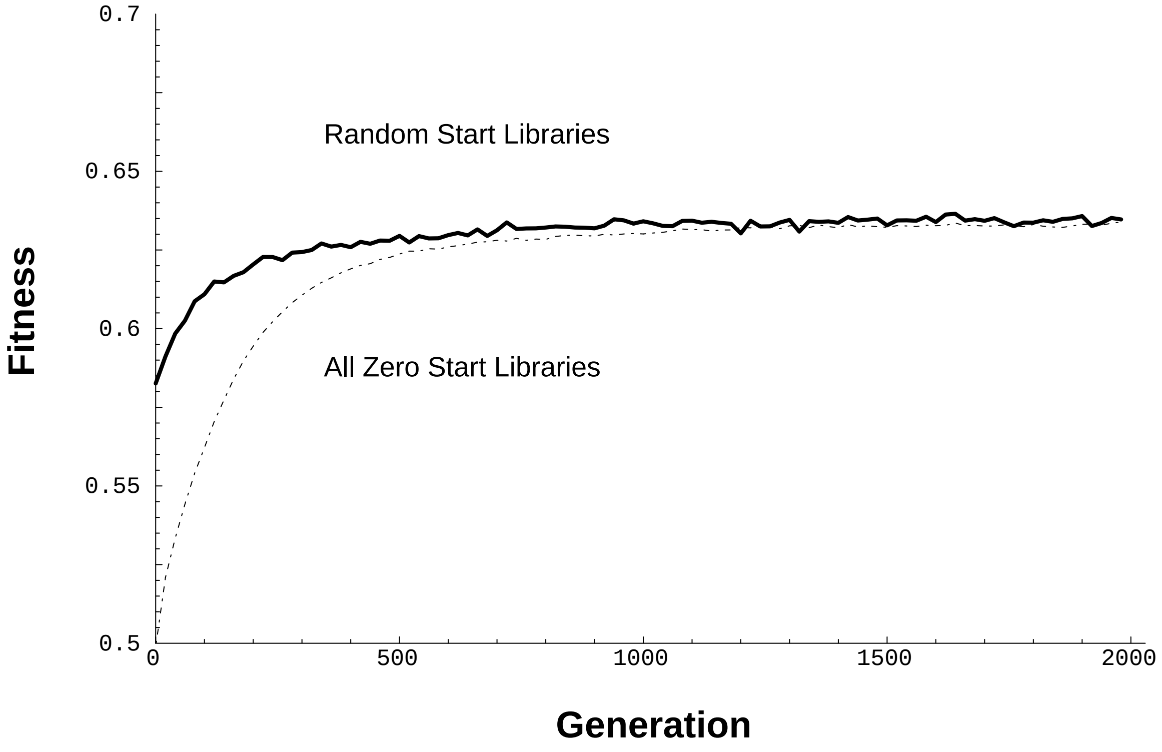

Figure 4: Evolution of gene libraries that initially are random. Here $U = 32, K = 8$, and $N = 8$.

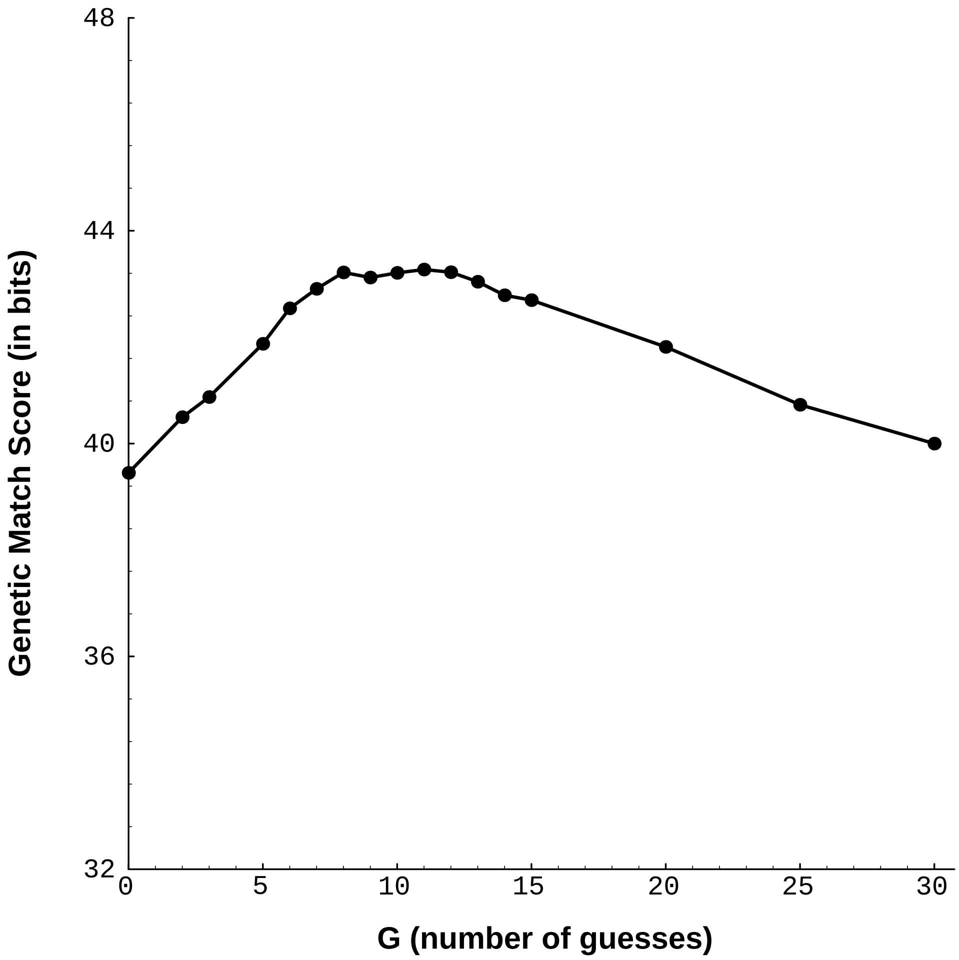

Figure 5: Somatic mutation can speed evolution. In this experiment the germline encoded antibodies in the expressed repertoire are allowed to improve their antigen recognizing ability by fixing some small number of the incorrect bits. The immune system expresses N antibodies. For each antigen seen, the antibody with the best match score is selected. Through somatic mutation this best antibody is allowed G trials in which to improve its match score. Each trial has a 50 percent chance of success, so on average the match score will improve by $G/2$. In the experiment we also introduce a threshold into the computation of a binding affinity. If the match score of the best matching antibody after mutation is below the threshold the affinity is zero; above the threshold the affinity is one. The affinity is the "score" for that antigen. This entire process of computing the "antigen score" is applied to each antigen and the total fitness of an individual is the average of the antigen scores. In this experiment a population of 50 individuals was followed for 1000 generations. The graph shows the average fitness at birth, i.e. the genetic component of the match score, for the population after 1000 generations computed as the average over 30 experiments.