Simulated evolution of antibody gene libraries under pathogen selection

Mihaela Oprea
Computer Science Department
University of New Mexico
Albuquerque
mihaela@cs.unm.edu

Stephanie Forrest
Computer Science Department
University of New Mexico
Albuquerque
forrest@cs.unm.edu

ABSTRACT

The immune system of vertebrates is generally viewed as a prototype of a highly adaptive, distributed, detection system, that identifies and neutralizes pathogenic intrusions. One of its puzzling features is that the immune receptors (antibodies) are able to bind to pathogens that they have not been "trained" to recognize. This anticipatory capability is thought to be due to a broad coverage of the pathogen space realized by the antibodies that the immune system can produce [3]. What we would like to understand is how this this coverage is achieved, given that the immune system uses a relatively small number of genes to construct its receptors. We use an evolutionary algorithm to explore the strategies that the antibody libraries may evolve in order to encode pathogen sets of various sizes. We derive a lower and an upper bound on the performance of the evolved antibody libraries as a function of their size and the length of the pathogen string. We also provide some insights in the strategy of the antibody libraries. We discuss the implications of our results for biological evolution of antibody libraries.

I. INTRODUCTION

The recognition of pathogens by antibodies is done in terms of intermolecular binding. Upon binding the pathogen, the immune system cells (B cells) start producing large amounts of antibodies that bind to the pathogens and facilitate their elimination by cells that can completely degrade them. The broad coverage of the pathogen space is thought to be ensured by the production of a large number of different antibodies that are assembled in a combinatorial fashion from a number of fragments [8]. The genes for these fragments reside in gene libraries, and are pasted together by a process termed rearrangement (Fig. 1). If the usage of fragments from the libraries was random, the number of different antibodies that the organism can make would be obtained by a simple multiplication of the sizes of the all the different libraries that contribute to one receptor.

There are a number of problem with this simple calculation. The contribution of the different gene fragments to the binding site of the antibody is different, with the V gene having the highest representation. Also, some antibodies that are very similar to each other contribute less to the diversity of the library. But the set of antibodies that we possess seems sufficient for the organism to withstand numerous infections over its life time. Thus, we expect the antibody gene libraries to somehow encode information about their pathogenic environment.

Hightower et al. [1] introduced an evolutionary algorithm that was used to study the performance of antibody libraries in a variety of circumstances, likely to be present in the biological evolution of antibody libraries [1, 7]. We use a similar model, with the intent of evaluating the strategy and performance of antibody libraries when confronted with pathogen sets of different sizes. Our evolved libraries use essentially two different strategies. When the pathogen set size is relatively small, it directly determines the structure of the antibody libraries. On the contrary, when the pathogen set is much larger than the number of available antibodies, the antibody library evolves to ensure maximal coverage of the complete pathogen set. Its structure becomes independent on the subset of pathogens that it evolved to match. We show that the transition between these two regimes occurs faster when the pathogens are allowed to mutate, and that in this case, the antibodies do not get to the optimal distribution in the sequence space.

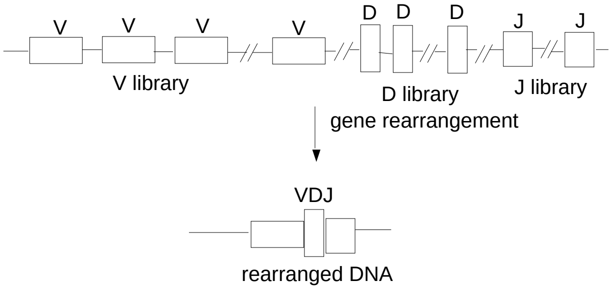

FIG. 1. The rearrangement process that leads to the formation of a functional antibody: the gene fragments (exactly one from each of the gene libraries) are concatenated in an orderly manner. The resulting product is then translated into the functional antibody molecule. V, D, and J are individual libraries that contribute to the production of the immune receptors.

II. BASIC MODEL

Let us introduce the components of our model.

- Assuming that different libraries contribute to the binding site of the antibody in an additive manner, we focus on the evolution of only one such library. We consider each individual having a genome of length $L \times A$ bits, composed of A individual genes. From this library, we assume that A antibodies are made, and that all these antibodies are available for binding any of the pathogens. Note that we do not distinguish between the genotype (antibody gene) and the phenotype (antibody molecule). We could alternatively view the libraries as representing the possible set of antibodies that an organism can produce. The genetic operators, to be discussed below, such as mutation and recombination on these libraries would then have to be thought to represent phenotypic changes to the antibody repertoire as a result of implicit genetic operations on the level of the genes.
- Pathogens are also represented as bit strings of length L.
- The essence of the complicated antibody-pathogen interaction in the real world, that we want to capture in our model, is that for each pathogen in the

environment of the organism there is at least one antibody in the individual's library that matches that pathogen. We use this property as the basis for our fitness function. To each individual library, $\mathcal{A}$, we assign a score in matching a pathogen p, which we define as

$$\sigma(p) = \frac{1}{L}(L - \min_{a \in \mathcal{A}} [h(a,p)])$$

where $h(a,p)$ is the Hamming distance between antibody a and pathogen p. In other words, for each pathogen, we find the antibody with the minimal Hamming distance to the pathogen. The score is a number between 0 and 1, being maximal for a perfect match, at Hamming distance 0, and minimal for the case of complementary bit strings. Note that we used identical lengths for the antibody and the pathogen strings and that we align the bit strings to calculate the Hamming distance.

- In Hightower et al.[1] the fitness f of an individual was identified with the average score $\langle\sigma\rangle$ with respect to all pathogens that it encounters. In this paper we will use the same fitness function. We believe that this choice can be most generally justified in terms of survival probabilities of an individual with respect to all pathogen challenges it encounters. For an organism to survive in a pathogenic environment, it has to meet successively all the pathogen challenges. Let us assume that the probability s_p to survive the attack of pathogen p grows exponentially with the score $\sigma(p)$. That is, for each additional matching bit between the best antibody and the pathogen, the probability s_p that the organism survives goes up by a constant factor, k. Thus,

$$s_p \propto k^{\sigma(p)}.$$

The probability to survive all pathogen attacks is given by the product of the survival probabilities s_p for all pathogens p. Therefore, the total survival probability s is given by

$$s \propto k^{P\langle\sigma\rangle}$$

where P is the number of pathogens, and $\langle\sigma\rangle$ is the score of the library averaged over all pathogens. Thus, we find that the survival probability s is a monotonically increasing function of the average score $\langle\sigma\rangle$. For the evolutionary dynamics defined on our antibody libraries, to be discussed later, only the relative ranking of the fitnesses of different libraries is important. Therefore, under the assumption that the fitness of an individual depends only on its survival probability s, we can identify the fitness with the average score $\langle\sigma\rangle$, without affecting the dynamics of the GA. Formally, if we denote the pathogen set by $\mathcal{P}$, the fitness f of an individual is given by

$$f = \frac{1}{P}\sum_{p \in \mathcal{P}} \sigma(p) \equiv \langle\sigma\rangle.$$

- In this paper we evolve the antibody libraries on a number of pathogen sets:
  - On the complete set of 2^L pathogens of length L.
  - On random subsets $\mathcal{P}$ of the complete pathogen set of size 2^L. These sets are constructed by sampling P pathogens, with replacement, from the complete pathogen set.
  - On pathogen sets that evolve independently of the individuals.
- Our genetic algorithm has the following structure:
  1. We choose a pathogen set $\mathcal{P}$ in one of the ways that we described above. We construct the initial population of $M = 50$ random libraries, of identical size, A. We found that, in conjunction with the parameters that we chose for the genetic operators, this value of the population size allowed convergence to a relatively high fitness solution. Also note that we start with a population of random libraries, as opposed to libraries composed of identical antibodies [2,7,1]. The motivation is that we intend to characterize the solutions with maximal fitness that the GA can find, regardless of the starting point.
  2. We determine the fitness of all libraries in the population with respect to the pathogen set, and we rank order the libraries according to fitness.
  3. We assign a weight $w_r = \frac{2(M-r)}{M(M-1)}$ to each individual, where r is the rank of the individual. Note that the weights sum to 1.
  4. To create one library of the new generation, we perform the following steps:
     (a) We select, with replacement, two libraries of the old population with probability equal to their weights. This selection scheme is called "rank selection" [6]
     (b) We generate two new libraries by crossing over the two chosen libraries. The number of crossover points n is chosen from a binomial distribution with mean $0.01A$. The crossover points are chosen at the boundary between antibodies, so individual antibodies are not disrupted by crossover.
     (c) We choose one of the two new libraries with equal probability.
     (d) We mutate it, with a probability of 0.002 per bit, and we add it to the new population.
  5. M new libraries are created using the previously described algorithm. The new population replaces the old population and we go to step 2 again.
  6. After a fixed number of generations, we determine the highest fitness library in the population, and we use this library to calculate various statistics.

III. RESULTS

A. The performance of a library of A antibodies on the complete pathogen set

Given that we start with random libraries, their performance gives us a lower bound on the fitness of evolved libraries. Let us determine the expected fitness of a random library on the complete pathogen set of size 2^L, where L is the length of the string representing the pathogen. Let $\sigma(p_i)$ be the score of an individual with respect to pathogen p_i and m the number of matching bit positions between a pathogen and an antibody. For a pathogen binding to a single random antibody, the probability that there are m or fewer matching bits, $Pr\{m \leq x\}$, is given by the value of the cumulative binomial distribution at x. If we have A antibodies, the probability that all of them have x or fewer matching bit positions with the pathogen is $[Pr\{m \leq x\}]^A$. Then the probability that the score $\sigma(p_i)$ of the individual with respect to pathogen p_i is x/L, is thus given by the probability that at least one antibody has x matching sites with the pathogen but none has more than x, i.e.

$$Pr\{\sigma(p_i) = x\} = [Pr\{m \leq x\}]^A - [Pr\{m \leq x-1\}]^A .$$

The expected score for pathogen p_i with a random antibody is then given by

$$E[\sigma(p_i)] = \frac{1}{L} \sum_{x=0}^{L} x Pr\{\sigma(p_i) = x\}.$$

The expected score of a random library on a random pathogen p_i also represents the expected score of a random library over the complete set of 2^L pathogens. We then denote the expected fitness of a random library over the complete pathogen set pathogens for a random library by f_r,

$$f_r = E[\sigma(p)].$$

The above equation for f_r gives a lower bound on the fitness of the evolved libraries as a function of L and A.

We can also calculate an upper bound for the fitness of the evolved libraries by using a theorem from the theory of error-correcting codes [4]. Assume that we distribute the A antibodies over the space of 2^L pathogen bit strings in such a way that each antibody a_i covers a set V_i of pathogens up to Hamming distance d. Assume that all sets V_i are disjoint and of equal size. In the best situation, there exists a Hamming distance d such that the sets V_i together exactly cover the space of 2^L pathogens. Since

$$V_i = \sum_{h=0}^{d} \binom{L}{h},$$

this yields the inequality

$$A \sum_{h=0}^{d} \binom{L}{h} \leq 2^L,$$

In the theory of error-correcting codes, this inequality is known as the sphere-packing or Hamming bound. The library is "perfect" if the equality holds. The fitness f_u of such a perfect library is given by

$$f_u = 1 - \frac{\sum_{i=0}^{d} i\binom{L}{i}}{\sum_{i=0}^{d} L\binom{L}{i}}.$$

Note that, for instance, given L and d, we can uniquely determine the library size A. Alternatively, given L and A we can determine the maximal value of d for which the inequality still holds.

How does the fitness of the evolved libraries compare to the bounds that we calculated? We used a string length $L = 8$ bits in order to test the scaling of the maximum fitness that the GA evolved with respect to the number of antibodies, A, in the library. A number of 10 GA runs were used for each data point in the figure. To test the dependence of the maximum fitness on the string length, L, given libraries of $A = 8$ antibodies, we used a hill climber approximation of the GA, since the runs of the GA algorithm become too computationally demanding. We previously tested the performance of the hill climber on this problem and we found it to converge to relatively good solutions, within less than 0.5% of the fitness value that the GA evolved. One run of the hill climber consists of the following steps:

1. start with a random antibody library of size A
2. calculate its fitness with respect to the complete pathogen set of size 2^L
3. mutate one bit of the antibody library
4. calculate the fitness of the new library
5. if this fitness is greater than or equal to the fitness of the old library, make this new library the current library, delete the old one, and go to step 2
6. if the fitness of the new library is smaller than that of the old library, discard the new library and go to step 3.

The total number of mutations that one library underwent was $(A \times L)^2$, A being the number of antibodies in the library and L the length of the antibody strings.

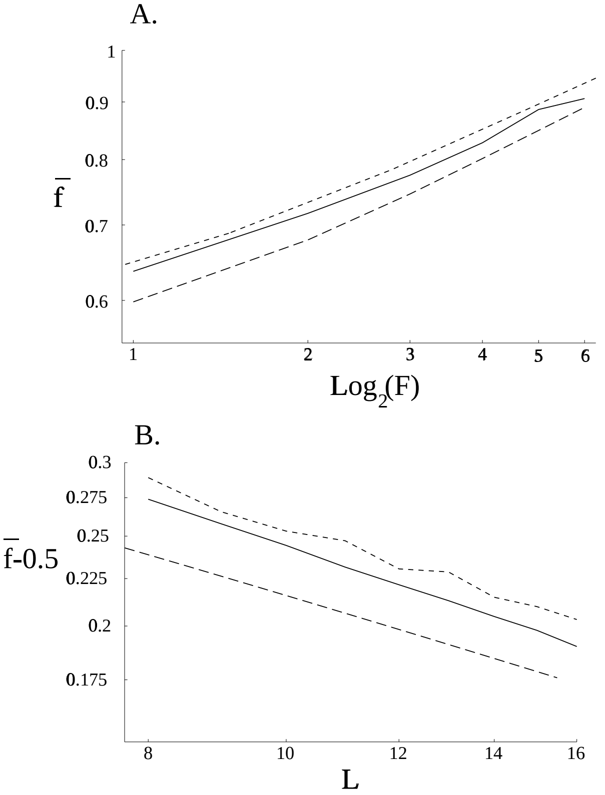

FIG. 2. Scaling of the fitness $\bar{f}$ with respect to the antibody set size A (panel A) and the string length L (panel B).

Fig. 2 panel A shows the best fitness $\bar{f}$ averaged over 10 runs of the GA, as a function of the *logarithm* base 2 of the number of antibodies A in the libraries as a solid line. The strings were length $L = 8$ and the fitness of the libraries was computed over the complete set of 2^L pathogens at each time step. The coarsely dashed line shows the lower bound f_r of the fitness of random libraries, while the finely dashed line shows the upper bound f_u obtained from the sphere-packing bound. This upper bound curve is obtained by plotting $f_u(d, L)$ against $\log_2[A(d, L)]$ for a range of values of the ball radii d, and $L = 8$. We plotted both the fitness, $\bar{f}$ and the logarithm $\log_2(A)$ on logarithmic scales, to illustrate the scaling relation that we inferred, namely that the the fitness of the libraries is proportional to a power of the logarithm of the number of antibodies A in the library,

$$f = c \log^{\alpha}(A).$$

For the results of figure 2A, the exponent is roughly given by $\alpha \approx 0.2$. The c values however differ between the different curves, and the basis for this difference will be explored in a later section. This scaling relation holds for the evolved libraries as well as for the upper and lower bounds f_u and f_r. However, the evolved libraries apparently do not manage to reach the fitness f_u as given by the "perfect" libraries. It is not clear to us, however, if this upper bound is realizable at all.

The dependency of the average best fitness on the length of the antibody and pathogen strings, L, is shown in Fig. 2 panel B. The y-axis shows $(\bar{f} - 0.5)$ on a logarithmic scale, while the x-axis shows L on a logarithmic scale as well. The solid line shows the best fitness $\bar{f}$ averaged over 10 runs of the hill climber as a function of the string length L for libraries of size $A = 8$. The fitness was computed over the full set of 2^L pathogens at each time step. The coarsely dashed line shows the lower bound f_r for the random libraries, while the finely dashed line shows the upper bound f_u as given by the sphere-packing bound. The upper bound is slightly more involved to determine in this case. Given the values of L and A, we can calculate the largest value of d for which the sphere-packing inequality still holds. This means that A disjoint spheres of radius d cover less than 2^L antibodies, but that A spheres of radius $(d+1)$ do not fit into the space of 2^L strings. We assume that the upper bound is obtained by distributing the A antibodies such that each of the antibodies covers a number $V(d) = \sum_{h=0}^{d} \binom{L}{h}$ pathogens at Hamming distance d or less, and that the remaining $2^L - A \times V(d)$ pathogens are partly shared between antibodies and are matched at a distance $(d+1)$. This leads to the upper bound

$$f_u = 1 - \frac{A}{2^L} \left(\sum_{i=0}^{d} \frac{i}{L} \binom{L}{i} \right) - \frac{d+1}{L} \left(1 - \frac{A \times V(d)}{2^L} \right).$$

The straight lines of figure 2B show that the evolved fitness $\bar{f}$ as well as the upper and lower bounds f_u and f_r obey the scaling relation

$$f = 0.5 + \frac{c}{\sqrt{L}}$$

Again, the fitness values of the evolved libraries nicely lie between our theoretical upper and lower bounds. Thus, it seems that the scaling of $\bar{f}$ as a function of L and A is similar to the scaling of the fitness of random libraries as well as "perfect" libraries. This suggests that these scaling relations are mostly a result of the geometry of the bit string space and the additive nature of the matching rule.

What implications do these scaling relations have? As 0.5 would be the fitness of the library in the limit of $L \to \infty$, $\bar{f} - 0.5$ represents the improvement in fitness that we can obtain, given that the pathogen (and antibody) strings have finite length. As the pathogen strings that the antibodies have to match become shorter, the fitness of a constant size library increases, but relatively slow, as the inverse of the square of the string length. The slower than logarithmic scaling of the fitness as a function of the number of antibodies A, means that if doubling the size of the library would lead to a fitness increase of δf, than doubling the size of the library *again* would lead to an increase of fitness smaller than δf. To obtain an increase of δf in fitness one would have to multiply the size of the libraries by larger and larger factors. A similar dependency was suggested, on experimental grounds, and within a somewhat different model, by Minar [5].

B. Performance of a library of A antibody genes as a function of the pathogen set size

Let us now turn to the case of the pathogen set being itself a subset of the set of strings of length L.

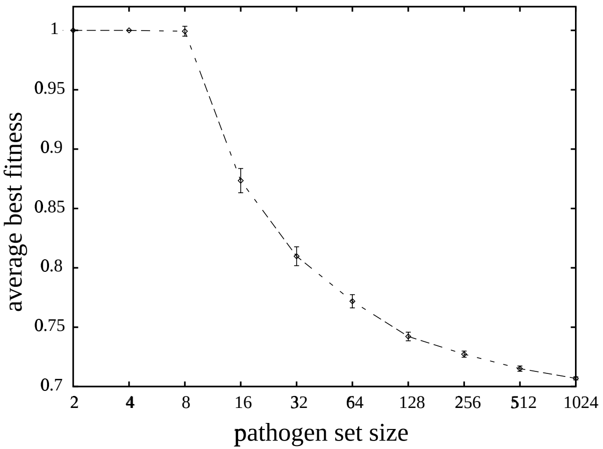

FIG. 3. Average fitness of the best individual evolved in a number of runs with different pathogen set sizes. 100 runs were used for each data point of pathogen set size 2, 4, 8, 16, 32, 64; 50 runs were used for the pathogen set size 128; 25 runs for pathogen set size 256, and 10 runs for pathogen set sizes 512 and 1024. Antibody library size $A = 8$. Length of antibody and pathogen strings, $L = 16$ bits. The errorbars indicate one standard deviation around the mean values.

Let us consider an antibody library of size $A = 8$ genes and study its performance as a function of the pathogen set size. In order to do this, we choose a pathogen string length of $L = 16$ bits, and we sample pathogen sets of sizes varying from $P = 2$ to $P = 1024$, with replacement, from the complete pathogen set of size 2^{16}. We perform a number of GA experiments for each pathogen set size, keeping the pathogen set constant during each run, and we store the best library evolved in each of these runs. Fig. 3 shows the results of these experiments. It is not surprising that the fitness deteriorates as the size of the pathogen set increases. The most dramatic effect is observed when the pathogen set size becomes larger then the size of the antibody library, as a perfect match cannot be ensured anymore for all the pathogens in the set. The strategy evolved by the libraries for covering a very large pathogen set is the subject of the next section.

C. The strategy of antibody libraries

What strategy then do the immune receptor libraries evolve in order to cover the pathogen set? We expect that for small pathogen sets, the antibodies would tend to match the pathogens perfectly. In this case the structure of the immune receptor library directly reflects the structure of the pathogen set. What we do not know is what strategies these libraries develop when they have to cover a pathogen set that is much larger than the size of the library. To address this question, we performed the following GA experiment. We used the libraries that we evolved previously on different pathogen set sizes. Consider one of these libraries, evolved to match a subset of P pathogens. We determine the mean, μ, and standard deviation, σ, of the fitness of this library on the complete set of 2^{16} pathogens. Then the average fitness of the library over a random pathogen set of size P is also μ, and the standard deviation on the set of P pathogens will be $\sigma/\sqrt{P}$. What we would like to know is whether the fitness of the library on the pathogen set that was used to evolve it is significantly different from the fitness of the same library on a random pathogen set of the same size. For this, we calculate the statistic

$$Z = \frac{f-\mu}{\sigma/\sqrt{P}},$$

where f is the fitness of the library on the pathogen set that was used to evolve it. The results, for library size $A = 8$, and string length $L = 16$, are plotted in Fig. 4. As we expect, when the pathogen set is of the same size as the antibody library, the libraries focus on these pathogens. Subsequently, if we are to test them on other pathogen sets of similar size, their performance is significantly lower. This behavior is maintained for a relatively large range of pathogen set sizes. As we get to large pathogen sets, the performance of the library on the pathogen set that it evolved to match becomes less and less distinguishable from the performance on a random pathogen set of the same size. Thus, as we vary the pathogen set size, the structure of antibody libraries changes, from from being completely determined by the pathogen set, to being independent of it. In this last regime, the antibodies in the library ensure a maximal coverage of the complete pathogen space.

Moreover, we can show that this transition occurs faster when we let the pathogens evolve as well. Let us slightly modify our genetic algorithm such that the pathogens also mutate at each generation of hosts. In one set of experiments we let each bit in each pathogen mutate with a probability of 0.1 bits per generation of the hosts. The actual number of mutations for each pathogen is chosen as a random deviate from the Poisson distribution with mean $0.1L$. The results are shown in the middle curve of Fig. 4. To test the effect of rapidly evolving pathogen set, we replaced a proportion of 1/8 of the pathogens at each host generation by random others. The results are shown in the lower curve of Fig. 4.

Thus, if pathogens evolve rapidly, or if indeed their number is much larger than the number of different antibodies that the organism can make, the structure of the pathogen set does not seem to get reflected in the antibody libraries. *Can we say anything about the structure of the antibody libraries in this situation?*

Choosing a string length $L = 9$ bits for antibodies and pathogens, we can test the libraries against the complete set of pathogens of this length. We evolved antibody libraries of size $A = 8$ to cover the complete set of pathogens of length $L = 9$, that is the pathogen set size was $P = 512$. The fitness of the best library was not the predicted value for a random library of size $A = 8$, but higher (0.760417 as opposed to 0.755414). It was previously conjectured that the antibodies evolve such as to maximize the average Hamming distance to the other antibodies in the library [1]. Let us determine the average pairwise Hamming distance between the antibodies in the library, and compare with the average pairwise Hamming distance in the libraries that we evolved.

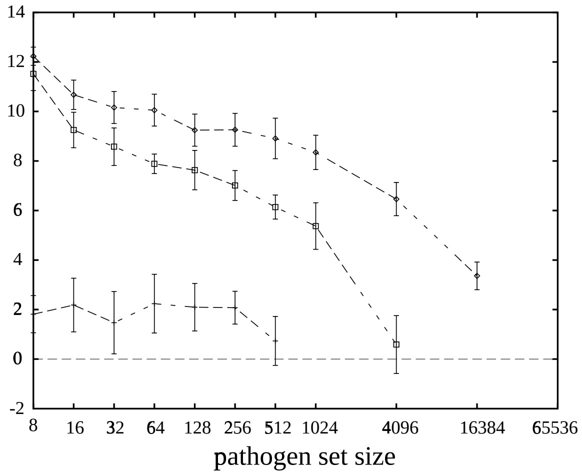

FIG. 4. Dependence of the Z statistic on the pathogen set size P. The size of the antibody library was kept constant, $A = 8$ genes. Length of antibody and pathogen strings is $L = 16$ bits. The three data sets are, from top to bottom, fixed pathogen set, slowly mutating pathogen set, rapidly evolving pathogen set. When the pathogen set that the library evolved to match was kept constant (upper curve) 100 runs were used for pathogen set sizes 2, 4, 8, 16, 32, 64; 50 runs for pathogens set sizes 128 and 256; 25 runs for pathogen set size 512; 10 runs for pathogen set size 1024, and 5 runs for pathogen set size 4096. For the evolving pathogen cases, 10 runs were performed for each pathogen set size, with the exception of the pathogen set size of 4096, for which 6 runs were used.

The expected pairwise Hamming distance within a library is given by

$$\frac{2}{A(A-1)}\sum_{i=1}^{A}\sum_{j=i+1}^{A} h(a_i,a_j)$$

where A is the number of antibodies in the library, a_i and a_j are individual antibodies, and $h(a_i, a_j)$ is the Hamming distance between them. This is given by

$$h(a_i,a_j) = \sum_{k=1}^{L}\delta(a_i^k,a_j^k)$$

where a_i^k and a_j^k denotes the k^{th} bit position of the two strings, and

$$\delta(a_i^k,a_j^k) = \begin{cases} 1 & \text{if } a_i^k \neq a_j^k \\ 0 & \text{otherwise} \end{cases}$$

We may now switch the order of summations to obtain:

$$\langle h\rangle = \frac{2}{A(A-1)}\sum_{k=1}^{L}\sum_{i=1}^{A}\sum_{j=i+1}^{A}\delta(a_i^k,a_j^k)$$

and since the bits are independent, maximizing this quantity means maximizing the pairwise Hamming distance at each bit position. If for a bit position k we denote by n_0 the frequency of 0's in the antibody population, then the pairwise Hamming distance at that position is $n_0(A-n_0)$, which is maximal for $n_0 = A/2$. Substituting in the above equation, we obtain for the maximal Hamming distance

$$\langle h \rangle = \frac{LA}{2(A-1)}$$

Coming back to the libraries that we evolved previously, with $A = 8$ genes, and string length $L = 9$ bits, the optimal Hamming distance is 5.143 bits. This Hamming distance was indeed found in the libraries that evolved the maximal fitness in this series of runs. This value is also significantly different from the average pairwise Hamming distance in random libraries. We tested this by generating a set of 10^4 random libraries, for which we determined the average pairwise Hamming distance. The value of our evolved library is different at a significance level $\alpha < 0.02$. Though having maximal average Hamming distance between the genes in the library seems to be a necessary condition for maximal fitness, it is not sufficient. Clearly, a library of size $A = 8$ composed of four copies of a string and four copies of its complement has maximal average pairwise Hamming distance, but it is far from being optimal. It is so far unclear what other condition needs to be fulfilled for a library to achieve maximal fitness. On the other hand, libraries that evolve in a rapidly changing pathogenic environment with a relatively small number of pathogens that select them, cannot be distinguished, by the average pairwise Hamming distance, from random libraries of the same size.

IV. IMPLICATIONS FOR ANTIBODY LIBRARIES ENCOUNTERED IN VARIOUS SPECIES

Let us now put in perspective the assumptions behind our model of antibody gene evolution, and summarize its predictions.

We considered that all hosts have an identical number of antibody gene libraries of constant length. We assumed that the strings representing the pathogens are aligned with the antibodies, and the score of one antibody with one pathogen is given by the proportion of positions at which the two strings match. We assumed that all the antibody types that an individual can make are available for interaction with every pathogen. The fitness of an individual is the score averaged over the whole set of pathogens.

The fitness that the evolved antibody libraries obey the same scaling laws as the theoretical lower and upper bounds that we derived. This suggests that the scaling of the fitness with library size and string length is not a function of the evolutionary dynamics, but is determined by the geometry of the bit space and the matching rule that we used. We can further analyze the implications of these scaling laws for the selective pressures that might operate in biological evolution of gene libraries.

The size of the binding site of antibodies is presumably under evolutionary pressure as well. With a given number of antibodies, the organism would achieve higher fitness by using short antibodies, that would bind to limited regions on the surface of the pathogens. The lower limit on the size of the recognition site is probably set in nature by the trade off between the quality of recognition and the specificity of it. That is, the shortest length of a string that allows differentiation between pathogenic motifs and motifs that are present in the proteins of the host.

We also deduced that the fitness of the antibody library increases only logarithmically with its size. In all the organisms in which extensive sequencing of the immune receptor locus has been performed, the number of genes that was found is of the order of a hundred. The recognition of pathogens is essentially a two stepped process. Somatic mutation, operating on the receptors that already bound the pathogen once, can improve their affinity by one to three orders of magnitude. The pathogen initially gets recognized by on of the germline-encoded receptors, after which the recognition of the pathogen is improved by somatic mutation of the germline receptors. It seems therefore, that there is an evolutionary trade off between increasing the size of the antibody library, versus improving the efficiency of the somatic mutation proces. The extremely slow increase in fitness with the size of the antibody libraries that we found in this study, raises the question of what mechanism would keep evolution from evolving smaller libraries of antibodies than the ones we actually observe. One possible explanation is that there is a recognition threshold in the matching between antibodies and pathogens below which recognition is not going to occur at all. In this case, some minimal number of antibodies would be required to ensure a complete coverage of the pathogen space. Alternatively, one may envisage the pathogen set structured as a distribution of clusters such that different genes in the library would reflect different clusters of pathogens. Furthermore, certain epidemies that have been caused in native populations by the migration of people from other geographical areas, suggest that individuals may indeed have different fitness in different pathogen environments. This suggests that pathogenic environments may indeed be relatively small and structured.

Acknowledgments

M.O. thanks Ron Hightower and Erik van Nimwegen for stimulating discussions. This work has been supported by the National Science Foundation Grant No. IRI-9711199 and the Office of Naval Research N00014-95-1-0364.

[1] R. Hightower. *Computational aspect of antibody gene families.* PhD thesis, University of New Mexico, 1996.

[2] R. Hightower, S. Forrest, and A. S. Perelson. The evolution of emergent organization in immune system gene libraries. In L. J. Eshelman, editor, *Proceedings of the Sixth International Conference on Genetic Algorithms*, Los Altos, CA, 1995. Morgan-Kaufmann.

[3] J. Kuby. *Immunology.* W.H. Freeman and Co., New York, 1992.

[4] F.J. MacWilliams and N.J.A. Sloane. *The theory of errorcorrecting codes.* Elsevier Science Publishers, B.V., 1986.

[5] N. Minar. Suboptimal solutions in a simple GA problem and the underuse of genetic material (unpublished manuscript). *http://www.santafe.edu/ nelson/gaimmune*, 1994.

[6] M. Mitchel. *An introduction to genetic algorithms.* The MIT Press, Cambridge, Massachusetts, 1996.

[7] A. Perelson, R. Hightower, and S. Forrest. Evolution (and learning) of Vregion genes. *Research in Immunology*, 147:202–208, 1996.

[8] S. Tonegawa, N. Hozumi, G. Matthyssens, and R. Schuller. Somatic changes in the content and the context of immunoglobulin genes. *Cold Spring Harbor Symposium on Quantitative Biology*, 41:877–888, 1975.