Using an Immune System Model to Explore Mate Selection in Genetic Algorithms

Chien-Feng Huang

Modeling, Algorithms, and Informatics Group (CCS-3),
Computer and Computational Sciences,
Los Alamos National Laboratory, MS B256,
Los Alamos, NM 87545, USA
cfhuang@lanl.gov

Abstract. When Genetic Algorithms (GAs) are employed in multimodal function optimization, engineering and machine learning, identifying multiple peaks and maintaining subpopulations of the search space are two central themes. In this paper, an immune system model is adopted to develop a framework for exploring the role of mate selection in GAs with respect to these two issues. The experimental results reported in the paper will shed more light into how mate selection schemes compare to traditional selection schemes. In particular, we show that dissimilar mating is beneficial in identifying multiple peaks, yet harmful in maintaining subpopulations of the search space.

1 Introduction

In the setting of multimodal function optimization, engineering and machine learning, there are two important issues when a GA is used: (1) how fast can a GA discover one or several peaks? And (2) can a GA maintain diverse subpopulations in different parts of the search space?[1] In this paper, we intend to use the mate-selection framework proposed in [7] and present the research work for investigating these two themes. In [7], it was shown that mate selection plays a crucial role in GA's search performance. In a nutshell, the dissimilarity-based mate selection schemes facilitate locating a single, best-so-far solution at the expense of generating lethal offspring; and the similarity-based mate selection schemes enhance selection pressure toward highly-fit individuals such that the GA's population converges rapidly to a certain region of a fitness landscape. As such, for the first question, we would expect the dissimilarity-based mate selection to improve the GA's search performance with respect to that metric. On the other hand, our empirical results so far have showed that simple GAs with the mate selection schemes are all subject to convergence (i.e., the simple

[1] The first issue was briefly discussed in [7]. For the second issue, there are some practical problems where maintaining subpopulations are critical. An example is the application of genetic approach to decentralized PI controller tuning for multivariable processes in [12].

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2723, pp. 1041–1052, 2003.

GAs cannot maintain subpopulations). Thus for the second question, we intend to employ Smith et al.'s immune system model [11], which was shown to be able to maintain diverse subpopulations, in order to offer additional insights into how the mate selection schemes compare to traditional selection schemes. In particular, we are interested in studying how different mate choices affect the capability of Smith et al's approach for maintaining subpopulations. Since it has been shown, in [7], that the dissimilar mating mechanisms are harmful in the sense of producing more useless hybrids, we expect that such mating preferences will reduce the proportions of individuals in subpopulations. If so, the next question would be to study if reducing the probability of dissimilar mating (or increasing the probability of similar mating) can improve the capability for maintaining subpopulations.

This paper presents the preliminary results we obtained while investigating the role of mate selection in the two issues discussed above. Before delving fully into this paper, however, it is important to briefly review Goldberg and Richardson's fitness sharing mechanism [3] that serves as an idealized approach for maintaining population diversity, and present Smith et al.'s immune system model to discuss how it implements a form of implicit fitness sharing so as to facilitate formation of subpopulations. We then summarize the relevant framework for studying mate selection proposed in [7]. Section 3 presents experimental results that answer the two questions mentioned above. Finally, this paper is concluded with the insights obtained for the mate selection schemes and future research lines.

2 Relevant Work in Prior GA Research

2.1 Fitness Sharing

Fitness sharing was an idea motivated by Holland's discussion [6] in which the number of individuals occupying a niche is limited to that niche's carrying capacity. Goldberg and Richardson [3] then introduced a fitness sharing mechanism that induces population diversity by penalizing individuals for the presence of similar individuals in the population. The technique they proposed was shown to be an effective method for maintaining subpopulations over several high-fitness regions of the search space. However, it has two serious limitations: (1) the peaks must be equidistant or nearly so, and (2) setting σ_s (a critical parameter in the fitness sharing scheme that represents a cutoff distance, beyond which no sharing will occur) requires knowledge about the number of peaks in the search space. These limitations arise from the fact that fitness sharing is defined explicitly.

To avoid the difficulty of appropriately choosing σ_s Smith, Forrest and Perelson [11] introduced an algorithm that does not require explicit construction of the sharing function. Their approach can *implicitly* achieve fitness sharing that discovers for itself how many peaks are in the search space (including the case of not equally spaced peaks), and allocate trials appropriately. The idea is to use the metaphor of biological immune systems which can maintain the diversity

needed for it to detect multiple antigens. Then the GA, combined with the immune system idea, effectively distributes the population over several high-fitness areas of the search space.

2.2 Binary Immune System Model

The immune system model considered in this paper is based on a model introduced by Farmer et al. [1], where both antigens and antibodies are represented by binary strings. It is a simplification from the real biology in which genes are specified by a four-letter nucleic acid alphabet and recognition between antibodies and antigens is based on their three-dimensional shapes and physical properties. However, this abstract model of binary strings is rich enough for exploring how a relatively small number of recognizers (the antibodies) can evolve to recognize a much larger number of different patterns (the antigens).

In this binary immune system model, recognition is evaluated through a string matching procedure. The antigens are considered fixed, and a population of N antibodies is evolved to recognize the antigens using a GA. For any set of antigens, the goal is to obtain an antibody *cover*—a set of antibodies such that each antigen is recognized by at least one antibody in the population. Maintaining diverse antibodies is crucial for obtaining a cover [11].

An antibody is said to match an antigen if their bit strings are complementary (maximally different). Since each antibody may have to match against several different antigens simultaneously, we do not require perfect bit-wise matching. Many possible match rules are plausible physiologically (See [10] for examples). The degree of match is quantified by a class of match score functions $M : \mathit{Antigen} \times \mathit{Antibody} \rightarrow \Re$. For instance, M can simply count the number of complementary bits or M can identify contiguous regions of complementary bitwise matches within the string.

Smith et al. [11] adopted a model in which a fixed set of antigens is given, and the antibodies are initialized either to be completely random (to see if the GA can learn the correct antibodies) or initially given the answer by setting the population to include the correct antibodies (to test the stability of the answer). Their mechanism for fitness scoring is as follows:

1. A single antigen is randomly selected from the antigen population.
2. From the population of N antibodies a randomly selected sample of size σ is taken without replacement.
3. For each antibody in the sample, match it against the selected antigen, determine the number of bits that match, and assign it a match score.
4. The antibody in the sample population with the highest match score is determined. Ties are broken at random.
5. The match score of the winning antibody is added to its fitness. The fitness of all other antibodies remains unchanged.
6. This process is repeated for C cycles (typically one to three times the number of antibodies).

In this scheme, since an antibody's fitness is increased only if it is the best matching antibody in the sample, the fitness values of antibodies are interdependent. In [11] Smith et al. showed analytically how this procedure implicitly embodies fitness sharing. Furthermore, Forrest et al. [2] reported that this scheme can maintain subpopulations of antibodies that cover a set of antigens.

2.3 Mate Selection Schemes

Based on the idea of "assortative mating" used in biology, [7] proposed a framework to investigate the role of mate selection in GA's search power.[2] Simply stated, the goal was to shed more light into how specific mate selection schemes compare to traditional selection schemes. In case of similar mating, similar individuals are chosen for mating; in case of dissimilar mating, dissimilar individuals will mate with each other. That is, the selection-for-mating step of a simple GA [9] is modified as:

During each mating event, a binary tournament selection[3]—with probability one the fitter of the two randomly sampled individuals is chosen—is run to pick out the first individual, then choosing the mate according to the following schemes:

Tournament Selection (TS): Run the binary tournament selection again to choose the mate.

Tournament Dissimilar Mating (TDM): Run the binary tournament selection two more times to choose two candidate partners; then the one more dissimilar to the first individual is selected for mating.

Tournament Similar Mating (TSM): Run the binary tournament selection two more times to choose two candidate partners; then the one more similar to the first individual is selected for mating.

Random Dissimilar Mating (RDM): Randomly choose two candidate partners; then the one more dissimilar to the first individual is selected for mating.

Random Similar Mating (RSM): Randomly choose two candidate partners; then the one more similar to the first individual is selected for mating.

We use the Hamming distance as the similarity metric. Notice that in the mate selection schemes above if the two candidates are of the same Hamming distance to the first individual, then one of them is randomly selected.

In the five approaches above, the first individual is always sampled by the regular tournament selection. For TDM and TSM, there are two ways to affect an individual's probability of being selected. The first results from the fitness evaluation explicitly defined by a given test function. The second is from the preference of each individual over other individuals that possess certain characteristics. The two sources complicate the probability of an individual being

[2] See [7] for a comprehensive literature review of the relating mate-selection work in prior GA research and a detailed discussion on why the framework was proposed.

[3] Tournament selection is employed here for low computational cost.

selected for actual mating. It is expected that tournament selection contributes more selection pressure toward highly-fit individuals, and the mate preference refines the searching for mates. As for RDM and RSM, the selection pressure is reduced by removing the tournament selection acting upon the candidate mates. The only source that affects the mate selection probability is precisely the mating preference, which exerts a selection pressure on the population based on genotype.

3 Experimental Results

To illustrate the effects of mate selection on the subpopulation-maintaining ability of Smith et al.'s immune system algorithm (we call it the diversity algorithm from here on), we use a simple example in which antigen populations cannot be matched by a single antibody type. Consider an antigen population that is composed of 50% $000\ldots000$ (all 0's) and 50% $111\ldots111$ (all 1's). In order for an antibody population to recognize these antigens, there would need to be some antibodies that are all 1's and others that are all 0's. Thus, a solution to this problem requires the GA to maintain two different solutions simultaneously. This is an example of a "multiple peaks" problem because there are two incompatible solutions that are maximally different. Typically, on multiple-peaks problems it is difficult for simple GAs to distribute the population over several peaks of a fitness landscape (two different subpopulations of antibodies that match two types of antigens, in this case). This is because the selection pressure in a simple standard GA usually entails strong convergence tendency to only one peak. Even without selection pressure, genetic drift due to sampling error can still lead the GA to converge on one of the peaks [4].

Forrest et al. [2] reported in their numerical experiments that the GA with the diversity algorithm can effectively avoid strong convergence to one peak and distribute the population over multiple peaks. As has been discussed in the beginning of this paper, we expect the mate selection schemes play an important role in maintaining subpopulations. In particular, our objective is to address the following questions concerning the capability of the GA, along with Smith's algorithm, for maintaining subpopulations:

- Can the GA with different mate selection schemes maintain stable subpopulations of antibodies for recognizing different antigens, or does it always converge on one peak? If it can maintain diverse subpopulations, then
- Is the proportion of antibodies in each subpopulation being affected by different mating preferences?[4]
- Do different mating preferences have influence on the discovery time of antigens?

In light of pattern-recognition, Forrest et al. [2] pointed out that the immune system needs to recognize bacteria partially on the basis of the existence of

[4] How many antibody representatives must be in the population for an antigen to be identified is critical. See [2] for a detailed discussion.

certain unusual molecules that are inherently different from human cells, since many bacteria have cell walls made from polymers that do not occur in humans. With this as motivation, we study the GA's ability to detect common patterns (building blocks) in the antigen population and adopt the building-block idea in [6] to calculate fitnesses of antibodies.

Table 1. Building blocks of antigens

b_1 = 11111**********; s_1 =10
b_2 = *****11111**********; s_2 =10
b_3 = **********11111*****; s_3 =10
b_4 = **********11111; s_4 =10
b_5 = 00000**********; s_5 =10
b_6 = *****00000**********; s_6 =10
b_7 = **********00000*****; s_7 =10
b_8 = **********00000; s_8 =10

Table 1 illustrates the building blocks of antigens $111\ldots1$ and $000\ldots0$ (string length is of 20 bits[5]). An antibody is said to match an antigen if its bit string is complementary to the antigen at certain building blocks. Specifically, the match score function M_b is to identify the building blocks for which an antibody matches an antigen, and then assign corresponding scores to that antibody. For example, given an antigen $111\ldots1$, an antibody with the first five and the last five bits being all 0's will receive score $s_1 + s_4 = 20$, since these ten bits are complementary to those of the antigen.

Smith et al. [11] considered two cases for the score calculation of antibodies—perfect match and partial match. In case of perfect match, an antibody receives a non-zero score only if it perfectly matches the antigen. In case of partial match, an antibody receives a non-zero score if it partially matches the antigen. In terms of the distance d_{ij} between antibody i and antigen j, partial match indicates the degree by which an antibody matches an antigen—i.e., the number of bits of an antibody that are complementary to the corresponding bits of an antigen. The degree of match determines the specificity of an antibody. For example, if $d_{ij} = 0$, the matching is completely specific (that is, the antibody must perfectly match the antigen), but if $d_{ij} \neq 0$, it is partially matched. The consequence of a partial matching rule is that there is a trade-off between the number of antibodies used and their specificity—as the specificity of antibodies increases, so does the number of antibodies required to achieve a certain level of detection [5].

For the scoring rule discussed in the building-block-based recognition problem, we can also expand its definition by allowing partial match. In other words,

[5] The small string length here serves well for illustrating the effect of the mate selection schemes. We current have some results for larger string lengths that are consistent with the results obtained for the small string length.

Table 2. Illustration of the immune-based GAs.

1. Randomly generate an initial population of n antibodies.
2. Evaluate antibodies' fitnesses by the six steps of the diversity algorithm.
3. Repeat until n offspring have been created. a. select a pair of parents for mating by particular selection schemes; b. apply crossover operator; c. apply mutation operator.
4. Reset all the new individuals' fitnesses to zero and replace the current population with the new population.
5. Go to Step 2 until terminating condition.

if an antibody matches an antigen at all the bits of a building block, it is a perfect building-block match; if not all the bits of that building block are required for matching, it constitutes a partial building-block match. Therefore, the prefect building-block match case is that an antibody scores if all of its bits at a building block are complementary to those of an antigen. On the other hand, a case for partial match could allow an antibody to score with only 80% bits (i.e., 4 bits in case of the building blocks shown in Table 1) of a building block at which it matches an antigen. The result of this flexible scoring is a smaller population size required to achieve a certain level of recognition performance. In this paper, we mostly concentrate on this latter case for calculating antibody scores. (In case of 100% building-block match, a few experiments conducted so far have shown similar qualitative results as the 80% building-block match case, but it requires much larger population sizes, i.e., much higher computational costs, to achieve similar levels of performance.)

3.1 Effects of Mate Selection on Maintaining Subpopulations

To address the questions mentioned in the beginning of this section we conduct a series of GA experiments using the diversity algorithm. The illustration of the immune-based GAs is shown in Table 2.[6] Our first objective is to investigate effects of mate selection on the diversity algorithm's subpopulation-maintaining

[6] Since in the diversity algorithm the match scores of winning antibodies are continuously accumulated, after each generation their fitness values can be large. Thus at step 4 of Table 2 we reset the fitnesses of the new population's individuals to zero after each generation to prevent fitnesses from unlimited increase.

ability. Unless stated otherwise, these experiments use an antibody population size of 100, crossover rate of 0.7, mutation rate of 0.005, and ran for 150 generations. The antigen population is 50% $000\ldots0$ and 50% $111\ldots1$, and both antigens and antibodies are binary strings of length 20. The number of samples, σ, is 10, which is 10% of the population size. We choose this value because Smith et al.'s analysis suggests that too small or too large a sample size cannot show fitness sharing's effect. In addition, as mentioned in the preceding section, the number of cycles (C) does not have a bearing on the antibodies' expected fitnesses, 100 cycles (i.e., population size) used for each generation turned out to serve well for displaying subpopulation-maintaining results. Thus the total function evaluations for each run are generations×cycles×sample size, which equal 150,000.

Fig. 1 illustrates the experimental results of the diversity algorithm (averaged over 50 runs), evolved by the GAs with TS, TDM, TSM, RDM and RSM.

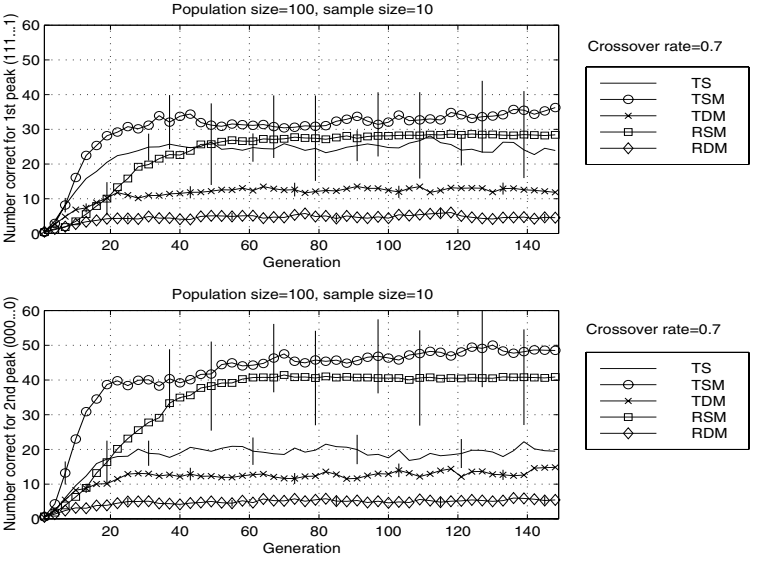

Fig. 1. The number of antibodies that correctly recognize antigens

These are the results for the numbers of antibodies that recognize antigens when all four building blocks are 80% correctly matched. Note that only the curves with small error bars (95% confidence intervals[7]) can be used for reliable judgements (we will discuss the reason for the larger error bars shortly), and thus the results for TS, TDM and RDM can be compared. It is clear that the

[7] The vertical bars overlaying the metric curves throughout this paper represent the 95-percent confidence intervals calculated from Student's t-statistic [8].

dissimilar mating schemes, TDM and RDM, generate less desired antibodies than the regular tournament selection. The reason is in the following:

When crossover is turned on (crossover rate is .7, in this case), the dissimilarity-based mate selection increases the probability of producing useless hybrids—e.g., given an individual $111\ldots1$, and two candidate mates $111\ldots1$ and $000\ldots0$, the GAs with the dissimilar mating schemes tend to choose $000\ldots0$ for mating with $111\ldots1$, and the crossing-over between these two strings generates offspring that fall into the valley between the two peaks. Therefore, TDM and RDM maintain a smaller fraction of desired antibodies.

On the other hand, we see that TDM generates a larger fraction of desired antibodies than RDM. The difference between theses two schemes is the method of selecting the second individual for mating—that is, in TDM fitter individuals have higher probabilities of being selected as mates, but this is not the case for RDM. As a result, TDM can pick out more individuals from the two peaks than RDM, which in turn increases the proportion of desired antibodies.

A remedy for the problem of producing useless hybrids would be to reduce dissimilar mating rates. In terms of the example above, the regular tournament selection confers $111\ldots1$ and $000\ldots0$ with equal probability of being selected for mating, thereby reducing the likelihood of two mating individuals chosen from the two peaks. However, if individuals tend to select similar mates, the selection pressure toward these individuals may be strong enough that the GA's population converges on only one peak. If this is the case, the diversity algorithm's capability for maintaining subpopulation is degraded.

The larger error bars for TSM and RSM in Fig. 1 illustrate this situation. Since TSM and RSM induce too strong a selection pressure, most of the GA's population members converge to only one peak. At generation 150, the GA with TSM has 20 (out of 50) runs in which most of the individuals converge to all 1's, and in 14 (out of 50) runs most of the individuals converge to all 0's, and there are 16 runs in which the two peaks are present, simultaneously. In case of RSM, there are 17 runs in which most of the individuals converge to all 1's, 21 runs in which most of the individuals converge to all 0's, and 12 runs where the two peaks are lost.

As a further illustration, Fig. 2 is the experimental results of a typical run for the number of desired antibodies obtained based on TSM. This figure shows that $000\ldots0$ are drown out by $111\ldots1$ after generation 60, although they do show up in earlier generations. This is because in TSM, similar individuals are always chosen as mates (with probability one)—a selection pressure toward similar mates enhances the convergence on one peak.

3.2 Effects of Mate Selection on the Discovery of Peaks

In the immune system problem considered, thus far we have been concerned with maintaining desired antibody subpopulations. However, there is another relevant issue we have not yet studied: the formation of the antibody subpopulations requires these antibodies to be discovered first. This is equivalent to the problem

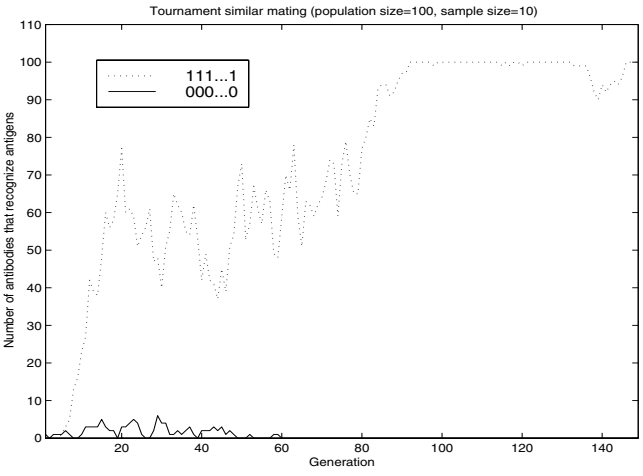

Fig. 2. The number of antibodies that correctly recognize antigens (based on the tournament similar mating), where all portion of the solid line (i.e., corresponding to $000\ldots0$) after generation 10 is on the 0 level

of finding multiple peaks. Since it has been shown, in [7], that the dissimilarity-based mate selection facilitates locating a single, best-so-far solution, we are interested in investigating if dissimilar mating is also more beneficial in finding multiple peaks than traditional selection schemes.

Table 3 displays the averaged mean function evaluations (over 50 runs) of discovering $111\ldots1$ and $000\ldots0$. These results show no obvious difference between various mate selection schemes for finding the two peaks, except that there are two runs where $000\ldots0$ was not found by the RSM GA, and this GA used a bit more evaluations to locate $111\ldots1$ than the other GAs. A closer inspection again shows the selection pressure toward similar individuals led the two particular runs of the GA to converge on $111\ldots1$, thereby precluding the discovery of the other peak. However, as population size decreases, the discrepancies between these mating schemes become more obvious. Table 4 illustrates the results for the number of runs (out of 50) in which antibodies $111\ldots1$ and $000\ldots0$ are discovered, respectively, based on population size 20 and sample size 2 (other parameter values remain unchanged). It is clear that the dissimilarity-based mating preferences facilitate locating two peaks. This is again because the similar mating schemes introduce a selection pressure strong enough that the corresponding GAs show inferior performance. All this confirms with our expectation that the dissimilarity-based mate selection is beneficial in locating multiple peaks.

Table 3. The mean function evaluations of discovering antibodies $111\ldots1$ and $000\ldots0$ (over 50 runs)

Antibody	TS	TDM	RDM	TSM	RSM
$111\ldots1$	2340 (368)	2460 (333)	2460 (272)	2440 (204)	3060 (338)
$000\ldots0$	2300 (206)	2540 (323)	2320 (270)	2180 (224)	48 runs reached

Table 4. The number of runs (out of 50) in which antibodies $111\ldots1$ and $000\ldots0$ are discovered

Antibody	TS	TDM	RDM	TSM	RSM
$111\ldots1$	20	34	40	18	23
$000\ldots0$	28	37	34	23	23

4 Conclusions and Future Work

In this paper, we have described Smith et al.'s immune system model in which subpopulations can be maintained through specific interactions among the strings. We have emphasized the performance of the GA in the binary immune system model, investigating how mate selection affects the GA's subpopulation-maintaining ability and the effects of mate selection on the discovery of multiple peaks. Both of these issues are important in the setting of multimodal function optimization, engineering and machine learning.

In studying the subpopulation-maintaining problem, the results illustrate that the dissimilar mating schemes are harmful in the sense of producing more lethal offspring. Consequently, the proportion of individuals that are representatives of different antibodies is reduced. We then showed that reducing the probability of dissimilar matings can remedy this problem. We also hoped to improve the GAs' performance by further increasing similar mating rates. However, as shown by the results obtained for TSM and RSM, they introduce a selection pressure strong enough that the population converges on only one peak.

In studying the peaks-identifying problem, we showed that the dissimilarity-based mate selection schemes facilitate locating multiple peaks of the fitness landscape. This is a crucial extension of the results obtained in [7], where dissimilar mating is shown to be more advantageous in finding a single, best-so-far solution.

Since the pattern-recognition strategy in our approach was based on schema detection, it is worth further exploration because in real problems when there are many more antigens than antibodies, antibodies need to detect common regions. In future work, we also hope to extend the results of schema detection and multiple-peaks identification to more realistic scale of antigens and antibodies. Finally, we would like to develop an analytical analysis to enhance our understanding for mate selection in the context of the immune-GA-based system.

Acknowledgments. The author would like to thank John Holland, Rick Riolo for their advice, and Bob Lindsay, Ted Belding, Leeann Fu, Tom Bersano-Begey and Bill Rand for their comments and suggestions.

References

1. Farmer, J. D., Packard, N. H., and Perelson, A. S.: The Immune System, Adaptation, and Machine Learning. In D. Farmer, A. Lapedes, N. Packard, and B. Wendroff (Eds.): Evolution, Games and Learning. North-Holland (1986). (Reprinted from Physica, 22D, 187–204)
2. Forrest, S., Javornik, B., Smith, R. E., and Perelson, A. S.: Using Genetic Algorithms to Explore Pattern Recognition in the Immune System. Evolutionary Computation, 1(3) (1993) 191–211.
3. Goldberg, D. E. and Richardson, J.: Genetic Algorithms with Sharing for Multimodal Function Optimization. Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms (1987) 41–49.
4. Goldberg, D. E. and Segrest, D.: Finite Markov Chain Analysis of Genetic Algorithms. International Conference on Genetic Algorithms, 2 (1987) 1–8.
5. Hofmeyr, S. A., and Forrest, S.: Architecture for an artificial immune system. Evolutionary Computation, 8(4) (2000) 443–473.
6. Holland, J. H.: Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press (1975).
7. Huang, C.-F.: A Study of Mate Selection in Genetic Algorithms. Doctoral dissertation. Ann Arbor, MI: University of Michigan, Electrical Engineering and Computer Science (2002).
8. Miller, R. G.: Beyond ANOVA, Basics of Applied Statistics. John Wiley and Sons (1986).
9. Mitchell, M.: An Introduction to Genetic Algorithms. Cambridge, MA: MIT Press (1996).
10. Perelson, A. S.: Immune network theory. Immunol. Rev., 110 (1989) 5–36.
11. Smith, R., Forrest, S., and Perelson, A. S.: Searching for Diverse, Cooperative Populations with Genetic Algorithms. Evolutionary Computation, 1(2) (1993) 127–149.
12. Vlachos, C., Williams, D., and Gomm, J. B.: Genetic Approach to Decentralized PI Controller Tuning for Multivariable Processes. IEE Proc. Control Theory and Applications, 146 (1999), 58–64.