On the Contribution of Gene Libraries to Artificial Immune Systems

Peter Spellward
Department of Computer Science
University of Bristol
Bristol, UK
+44 781 218 4941
Peter.Spellward@btinternet.com

Tim Kovacs
Department of Computer Science
University of Bristol
Bristol, UK
+44 117 954 5145
Tim.Kovacs@bristol.ac.uk

ABSTRACT
Gene libraries have been added to Artificial Immune Systems in analogy to biological immune systems, but to date no careful study of their effect has been made. This work investigates the contribution of gene libraries to Artificial Immune Systems by reproducing and extending an earlier system that used gene libraries. Performance on a job-shop scheduling problem is evaluated empirically with and without gene libraries, and with many different library configurations. We propose that gene libraries encourage diversity in a population of solutions and that the number of components in the gene library parameterises this effect. The number of gene libraries used is found to affect solution fitness and indeed using larger numbers of libraries (and therefore libraries of smaller components) enables higher fitness to be attained. We conclude that gene libraries are likely to be of use in applications where there is a need to maintain the diversity of solutions.

Categories and Subject Descriptors
I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search – *heuristic methods, scheduling*.

General Terms
Algorithms, Performance, Design, Experimentation, Theory

Keywords
Artificial Immune Systems, Gene Libraries, Job Shop Scheduling, Solution Diversity

1. INTRODUCTION
Artificial Immune Systems (*AIS*) (see [1, 2]) are computational systems that use metaphors and processes derived from the biological immune system [3]. Typically these are taken from the adaptive immune system; one of a number of layers that defend the body against invading organisms and toxins (*pathogens*).

GECCO'05, June 25-29, 2005, Washington, DC, USA.

Once pathogens enter the body, the adaptive immune system attempts to identify and eradicate them. This is not a simple task given the limited resources of the immune system and the vast and varying pathogens that it constantly has to defend against. The identification of foreign molecules (*antigen*) is performed by *lymphocyte* white blood cells that circulate around the blood and lymph systems. *Antibody* regions of the lymphocytes form a chemical bond to molecules they encounter and if the bond is sufficiently strong then the lymphocyte is likely to have encountered an antigen. An immune response is then induced which leads to the destruction of the pathogens.

Lymphocyte antibodies are produced through a pseudo-random combinational process utilising a set of inherited gene libraries [4]. This enables the body to maintain a diverse set of antigen detectors as the antibodies produced on one lymphocyte are likely to be different from those on any other lymphocyte. The body's lymphocyte population is also dynamic as lymphocytes normally live for just a few days before being replaced and so the new population identifies a constantly changing subset of antigens. In [1] it was estimated that every 10 days there is a completely new repertoire of lymphocytes in the human immune system. These two mechanisms ensure the body's adaptive immune system defence is kept dynamic and diverse, and maximises the effectiveness of its resources. Although humans and mice have fewer than 10^5 genes in their entire genome, through the use of gene libraries the immune systems of these species have been found to produce approximately 10^{11} different antibodies ([5, 6]).

Hart et al. [7] applied this biological analogy to the shop-floor environment to produce schedules for a job-shop scheduling problem (*JSSP*) that were robust to foreseeable and unforeseeable delays. In the main, work on JSSPs has concentrated on producing systems that generate optimal schedules (see [8, 9]), which attempt to minimise some criterion such as make-span or job tardiness. In the real world however optimal schedules can be extremely fragile; a slight delay caused by a machine breakdown for example may render the schedule extremely sub-optimal and require the costly process of rescheduling to take place ([10, 11]). Uniquely Hart et al.'s system employed a genetic algorithm to evolve populations of AIS capable of generating reasonable schedules robust to possible delays and so avoiding or considerably reducing the rescheduling problem.

Hart et al. focussed on a benchmark JSSP from [12] of 15 jobs (j) and 5 machines (m). Each job of a JSSP had an associated arrival date and due date, and consisted of a number of operations each of which required a different machine for a fixed period of time.

The fitness of a schedule was defined as the tardiness of the latest operation to finish after the job's due date.

In Hart et al. antibodies were considered to indirectly represent schedules and antigens to be synonymous with possible delays. AIS gene libraries were evolved to produce antibodies, and therefore schedules, aimed at minimising job-tardiness on a training set of antigens (foreseeable delays). An immune response could then be induced to rapidly generate good schedules from an AIS to combat new antigens (unforeseen delays).

Typical AIS implementations do not evolve gene libraries that then produce antibodies but rather directly evolve the antibodies. This latter approach is simpler, and there has been to date no principled study of what characteristics gene libraries can contribute to an AIS application.

Although Hart et al.'s system did make use of gene libraries, no experiments were reported comparing alternative library configurations or indeed what gene libraries contributed to the system.

In this investigation Hart el al.'s system was reproduced and extended to use configurable gene libraries. Significant experimentation was performed with various configurations to ascertain the effects of varying the number and size of gene libraries and to identify exactly what gene libraries can contribute to AIS applications.

2. SYSTEM OVERVIEW

Gene libraries are coded in the DNA of biological organisms and as a result are heritable and provide an initialisation bias to offspring. A single gene library consists of a number of components or gene fragments (c) of equal size (s). Antibodies are generated from a number of gene libraries (l) by randomly selecting a fragment from each library in a defined order and concatenating them.

This structure and process was simulated by Hart et al. and therefore in the replicated system. Individual AIS were defined as having 5 gene libraries ($l=5$) containing 5 components ($c=5$) of 15 elements ($s=15$) and so producing antibody representations of length jm. An AIS therefore contained lcs elements and could generate up to cl antibodies of ls elements. This was denoted the *potential antibody repertoire*. Figure 1 shows a simplified version of the combination process using 3 libraries of 3 components, with fragments of length 3 ($l=3, c=3, s=3$).

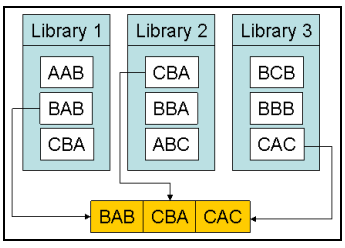


Figure 1. An example of how an antibody is generated from 3 gene libraries ($l=c=s=3$).

An antibody was used to indirectly represent a schedule using a method devised by Fang et al. in [10]. Each antibody representation consisted of a sequence of letters where each letter related to a job, allowing JSSPs of up to 26 jobs to be modelled. The order of the letters was used to define which unscheduled job task should be sequenced next. To ensure that a solution could be modelled in which all of the j jobs were completed using the m machines, the antibody was of length jm with each job appearing an equal number of times. For example, the antibody constructed in figure 1 is valid for a 3 job and 3 machine JSSP as it contains 3 occurrences of jobs A, B and C, one for each machine, and the constraint $jm=ls$ is satisfied.

An antigen was used to describe a set of arrival dates for the jobs to commence. The arrival dates were (with probability Pu) tardy variations of each job's expected arrival date and could be up to 300 units of time later than the original, subject to the constraint that the new arrival date allowed sufficient time for all of the job's tasks to complete (processing time required) before the due date. Antigens therefore represented a set of possible delays. During training an AIS could be exposed to 10 generated antigens from the benchmark 5x15 JSSP with $Pu = 0.2$, which formed the *Antigen Universe (AU)*.

Because of the computational costs involved only a subset of an individual AIS's potential antibody repertoire was expressed (N antibodies). This was denoted the *expressed antibody repertoire* and corresponds to the set of schedules an AIS can express. Additionally, each antibody expressed was only exposed to a subset of the AU (K antigens, corresponding to K variations on the expected arrival dates.). As the fitness for an AIS was based on a sample of its potential repertoire (i.e. its expressed repertoire) against a sample of the AU, an individual's fitness is based on an incomplete sampling of the environment. Indeed, selection pressure using this method only operates on the phenotype (the schedules produced). From investigations such as [13], however, it has been shown that this is sufficient to drive changes in the genotype.

The ability of an antibody to create a schedule that met the fixed due dates of the JSSP given the arrival dates of an antigen determined its *MatchScore*. The MatchScore was taken to be the tardiness of the latest job to finish after its due date (T_{max}). A match score of zero therefore indicated an antibody produced a schedule that enabled all of the jobs to be completed by their due dates. To define this more precisely; if each job j has a due date of Dj and completes at time Cj then the maximum tardiness T_{max} of a job was:

$$MatchScore = T_{\max} = \max\left(0, C_j - D_j\right)$$

Figure 2 gives an illustration of how the MatchScore was calculated for an antibody given a JSSP and an antigen selected from an AU.

The ability of an AIS to produce quality schedules for a particular antigen was denoted the *AntigenScore*. This was simply calculated as the lowest MatchScore found from a set of expressed antibodies on a single antigen:

$$AntigenScore = \min(T_{\max})$$

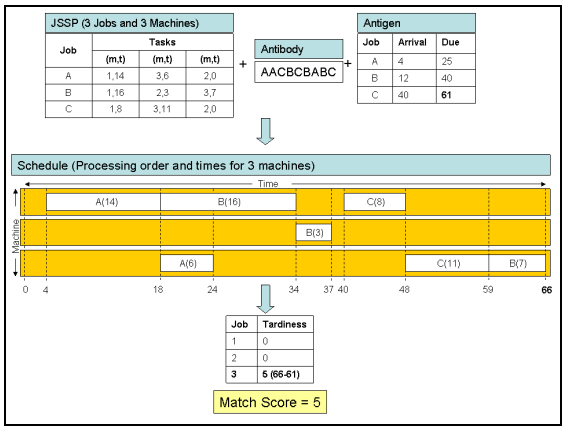


Figure 2. An illustration of how a schedule was derived from a JSSP, an antibody and the antigens it is exposed to. The match score is calculated from this schedule as the tardiness of the latest job.

To adapt the gene libraries, populations of individual AIS were evolved using a genetic algorithm, in which an individual's chromosome codes its gene library. The genetic algorithm used tournament selection, as well as probabilistic uniform crossover and mutation operators. Based on the findings of [2, 14, 15] the initial population was randomly generated to aid the rate of fitness increase (which we might call the learning rate). The fitness of each individual AIS in a population was calculated using the algorithm:

For each AIS

Express *N* antibodies at random from the AIS

Select *K* antigens at random with replacement from AU

For each of the *K* antigens

Calculate the MatchScores for the *N* antibodies

Assign the antigen the AntigenScore

End For

AIS fitness level = the average AntigenScore

End For

A low fitness (tardiness) value therefore indicated an individual had a large probability of being selected for proliferation into the next generation. The fitness of a population was measured as the average fitness of the AIS it contained.

The *Hamming Distance* of an AIS was computed by comparing the best schedules it produces (in response to an exposed antigen) pair wise, counting the number of places they differed and averaging the results. The Hamming distance for the population was taken as the average of the individual distance, and used as an indication of its diversity.

The same 5x15 JSSP (jb11) from [12] focused on by Hart et al. was used for all experimentation detailed in this paper.

Following [7] we have scaled fitnesses to lie in the range 0 to 1, where a higher value indicates a fitter individual. Optimal schedules therefore have a fitness value of 1. The exact scaling method used by Hart et al. was not detailed and could not be obtained. The following formula was therefore used to calculate the adjusted fitness (a) from the individual's (i) raw fitness (r).

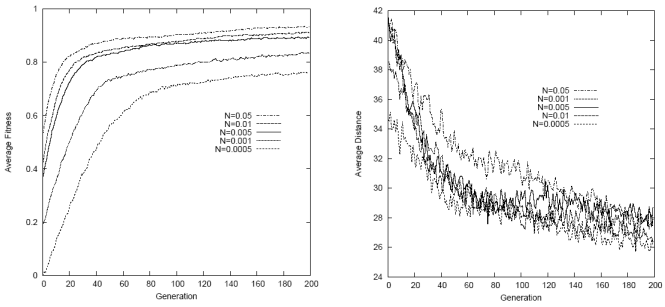
$$a(i) = 1 - \frac{r(i)}{r_{\max}}$$

In graphs, curve labels are ordered by the corresponding curve's final value in the accompanying key.

3. MODEL VERIFICATION

The system developed by Hart et al. was based on the AIS model described by Hightower et al. in [14] but extended to use gene libraries. Hart et al. replicated some of Hightower et al.'s experiments, and we replicated these same experiments to ensure the systems exhibited similar characteristics. Populations of 100 AIS *(l=5, c=5, s=15)* were evolved over 200 generations on the 5x15 JSSP with an AU of 10 antigens generated with *Pu=0.2*. Each experiment was repeated 10 times, and the results averaged.

In experiments varying antibody expression rates (N) using a constant antigen exposure level ($K=2$), our system demonstrated the same characteristics found by the two earlier studies. Specifically, by using higher antibody exposure rates higher fitness was attainable (figure 3a). It was further found that increasing the antibody exposure rates however caused the number of fitness evaluations to increase linearly. The diversity of the schedules was largely unaffected by the number of antibodies expressed. Through evolution, the diversity of the populations reduced as fitness increased indicating that the antibodies produced became more effective and specialised in combating the antigens in the AU (figure 3b).



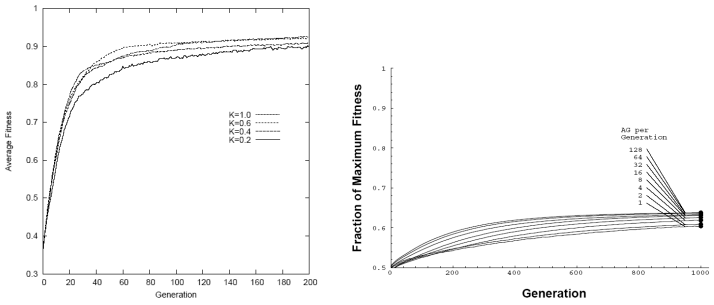
(a) **(b)**

Figure 3. (a) Fitness when varying antibody expression rate *N (%)*, with constant antigen exposure *K=2*. (b) Corresponding Hamming distance.

To this point, all three systems behaved in the same way. However, when antigen exposure levels (K) were varied while the antibody expression rate was static ($N=15$) Hart et al.'s results contradicted those of Hightower et al. Hart et al. found that increasing antigen expression rates reduced fitness while Hightower et al. found the opposite. Results of our replication

agreed with those of Hightower et al.: increasing antigen expression rates increased fitness. This can be seen in figure 4.

Notably we found that increasing antigen exposure levels also linearly increased the number of fitness evaluations required. Furthermore this study revealed that AIS evolved at lower levels of *K* exhibited higher amounts of diversity. We attributed this to the lower fitness, as the AIS were not able to specialise sufficiently in the 200 generations. Oprea and Forrest in [2] further hypothesised that other sources were more likely to play a greater contribution to antibody diversity than the antigens an individual is exposed to during its lifetime.



(a) **(b)**

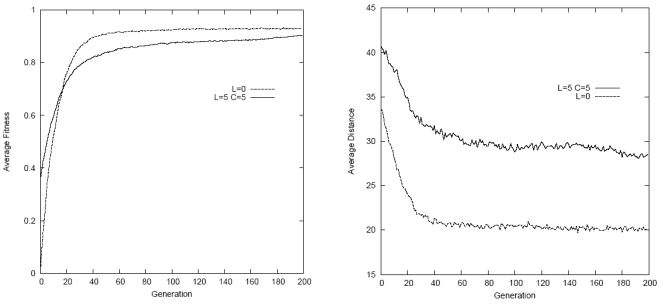
Figure 4. (a) Fitness for varying antigen exposure rates *K* (%) of an antigen universe of 10 antigens, at a constant antibody expression rate *N=15 (0.005%)*. (b) Results from [14] for a similar experiment where antigen exposure was varied between 1 and 128 antigen.

From our replication of earlier experiments we concluded that our system was consistent with the model of Hightower et al.

4. EFFECTS OF GENE LIBRARIES

To understand the role of gene libraries and what they contributed to the replicated application, comparisons were drawn between a version that used gene libraries and versions that did not.

By removing gene libraries each individual AIS in the population contained a single antibody and so therefore required a single fitness evaluation per antigen. In contrast, using the benchmark gene library configuration (*l=5, c=5, s=15*) and an expression level of *N=15* (0.005% of the potential repertoire), 15 evaluations were required for each antigen exposed to an individual AIS. We compared runs between the two configurations with populations of 100 individuals, an antigen exposure rate of *K=2* (0.2% of the AU), and averaged over 10 runs. Despite the large difference in the number of evaluations required the surprising results shown in figure 5 were produced.



(a) **(b)**

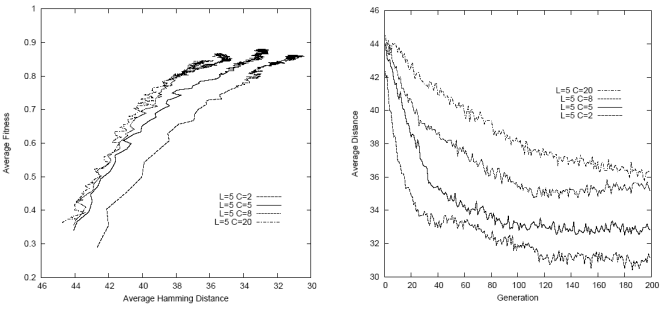
Figure 5. (a) Fitness (population average) results comparing a system with (*L=5*) and without (*L=0*) gene libraries. (b) The corresponding Hamming distance.

It was apparent that even though considerably fewer evaluations (800,000 compared to 12,000,000) were performed by the version without gene libraries it was able to generate fitter populations that produced schedules of far greater similarity. These experiments showed that gene libraries encourage diversity at a cost in fitness.

5. DIVERSITY INVESTIGATION

To identify which mechanism in gene libraries was the primary diversity contributor, experiments were performed on various gene library configurations. For all of the remaining experiments detailed in this report AUs generated with *Pu = 0.4* were used, the antigen exposure rate was set to *K= 4* (*0.4%* of the potential repertoire), and results were averaged over 10 runs. By altering *Pu* and *K* in this way we hoped to amplify any identifiable trends.

Initially we investigated the effects of varying component quantities. 5 gene libraries were used for all configurations while the number of components *c* was varied. *N* was set to expose 15 antibodies as per usual.



(a) **(b)**

Figure 6. (a) Comparison of Hamming distance and fitness of AIS with different numbers of components (b) Hamming distance of the schedules produced.

Although varying the number of components *c* had little effect on the ultimate fitness or the rate at which it was reached (figure 6a) it impacted the diversity of the generated schedules (figure 6b). It appeared that increasing the number of components meant that

more diversity could be maintained for longer i.e. increasing the diversity. The effect was quite dramatic and so the quantity of components used in gene libraries was identified as potentially being the main diversity contributor in AIS. Altering the number of components however also affected the size of the potential repertoire as does altering the numbers of gene libraries used. Further experiments were therefore carried out to ascertain if it was the size of the potential repertoires that affected diversity.

As the potential repertoire could be defined by c^l, the configurations detailed in tables 1 and 2 were used to test potential repertoires of 32,768 and 14,348,907 different antibodies. Both sets of experiments produced the same trends. The graphs shown in figure 7 were generated from the 32,768 antibody experiments.

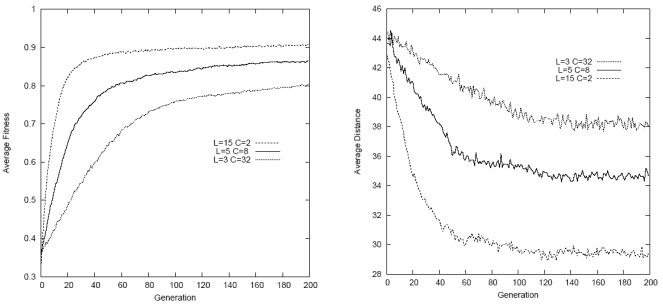
Tables 1 and 2. Gene library configurations required for creating potential repertoires of 32,768 and 14,348,907 antibodies.

Configurations for a Potential Repertoire of 14,348,907 Antibodies			
Libraries	3	5	15
Components	243	27	3

Configurations for a Potential Repertoire of 32,768 Antibodies			
Libraries	3	5	15
Components	32	8	2

The slope of the fitness curves (such as figure 7a) appeared to increase with the number of libraries used and to plateau at a higher level. This is investigated further in section 5. Comparing the fitness curves (figure 5a for the 32,768 antibody experiment) between the two sets of experiments it was found that higher fitness was attained with the smaller potential repertoire. This compliments the findings of [2]. We hypothesise that smaller repertoires achieve more rapid increases in fitness when the expression level is constant because a larger proportion of the potential repertoire is expressed.

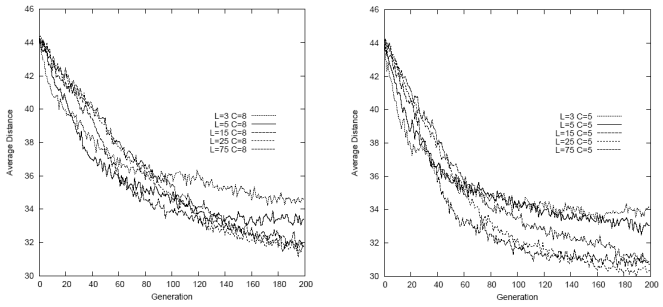
It was evident from the diversity results in figure 7b that although each configuration had the same size potential repertoire the distance trajectories differed. This suggested that it was not the size of the potential repertoire that affected the diversity as similar trajectories would have been found if that were the case. Of course, to generate differently configured AIS that had the same sized potential repertoire it was necessary to change the number of libraries used. To determine whether altering the number of libraries had affected the results experiments were carried out on configurations that contained the same number of components but different numbers of gene libraries.



(a) **(b)**

Figure 7. (a) Fitness of AIS with the same size potential repertoire (32,768). (b) The corresponding Hamming distance of the schedules produced.

To produce valid schedules the antibody produced had to be of length *jm* (in this case 75) and so the gene library constraint *ls = jm* had to be observed. Only library quantities that were factors of *jm* could therefore produce valid results. The diversity results for varying numbers of libraries of 5 and 8 components are shown in figure 8.



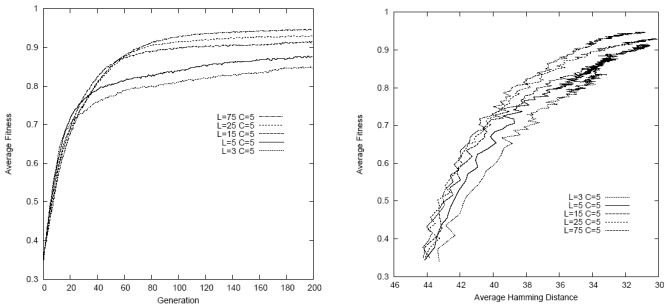
(a) **(b)**

Figure 8. Hamming distance for experiments using varying number of libraries consisting of (a) 5 components and (b) 8 components in each gene library.

Our results show that the number of libraries used had only a marginal effect on diversity and so could not have caused the large diversity changes found previously. The number of components contained in the gene libraries was hence determined to be the primary diversity mechanism of gene libraries with no significant detrimental effect on the rate of fitness development. No further computational cost is incurred either as the number of antibody-antigen encounters, and therefore fitness evaluations, remains static.

6. FITNESS INVESTIGATIONS

It had been hypothesised in section 5 that the number of gene libraries used affected the fitness of the population. To investigate this, experiments were carried out on AIS systems using varying numbers of gene libraries consisting of 5 components, subject to the constraint *ls = jm.*



(a) **(b)**

Figure 9. (a) The fitness of AIS with varying quantities of libraries and static number of components (b) The Hamming distance and fitness of the schedules produced.

The results (figure 9) showed that using more libraries enabled a higher ultimate fitness to be attained and indeed improved the rate at which it was reached. These results are supported by [2] and [16] which hypothesised that using more gene libraries could increase the rate of fitness increase as the size of the blocks (components) to optimise were smaller and therefore easier to optimise. The findings of these experiments support this hypothesis. A slight computational overhead is notable when increasing the number of gene libraries, as more crossovers are required during proliferation.

7. CONCLUDING SUMMARY

We found that gene libraries enabled significantly higher levels of diversity to be maintained although at the expense of reduced fitness. In investigating the mechanics of gene libraries it was further found that altering the number of library components used enabled diversity to be controlled. A greater number of components lead to higher diversity incurring only a marginal reduction in fitness. AIS that used more gene libraries were found to evolve to higher fitness more quickly with a marginal negative effect on diversity.

Previous studies [14, 17] hypothesised that species evolve gene libraries to respond best to the pathogens individuals are likely to face during their lifetime and that if the pathogen set is sufficiently large gene libraries evolve to generate antibodies that maximally cover the pathogen set. In our opinion, antigen coverage is maintained to prevent the situation where a specific set of pathogens is effectively targeted but the species becomes vulnerable to a large set of other pathogens. Our experiments further suggest that coverage is influenced by the diversity of the antigen produced. We might therefore suspect that species constantly under attack from a large number of pathogens will have gene libraries with a greater number of components than species exposed to relatively few pathogens

In this study gene libraries of uniform length and equal size components were used. In nature this is not necessarily the case as often gene libraries contain different numbers of gene fragments and indeed the size of gene fragments can differ from library to library. Further studies could simulate this to investigate when jagged libraries may be beneficial. Following our findings we hypothesise that smaller libraries would develop more specific areas of the antibody than larger libraries, subject to the antibody representation.

A number of minor factors were found to influence the fitness and diversity of the schedules produced e.g. antigen exposure rates. Further investigation could identify the bounds of these influences and indeed suggest when parameters ought to be altered with regard to the costs incurred.

In our implementation standard genetic operators were used by the genetic algorithm i.e. uniform crossover and mutation. Specialised genetic operators could be developed to improve the likelihood of producing quality antibodies and schedules. Examples of specialised recombination operators can be found in [18] and [19].

Gene libraries can be thought of as storing fragments of good antibodies evolved over generations. To make use of the fragments to solve problems such as rescheduling it is possible that combing immune system techniques with other methods such as case based reasoning could produce fruitful results.

8. ACKNOWLEDGEMENTS

The authors wish to thank from Steve Cayzer at Hewlett Packard Laboratories who suggested investigating the role of gene libraries in AIS, and also Emma Hart of Napier University, Edinburgh, who provided further information relating to the system described in [7].

9. REFERENCES

[1] Hofmeyr, S. and Forrest, S., *Architecture of an Artificial Immune System.* Evolutionary Computing, 2000. 8(4): p. 443-473.

[2] Oprea, M. and Forrest, S., *How the Immune System Generates Diversity: Pathogen Space Coverage with Random and Evolved Antibody Libraries*. In *Proceedings of the Genetic and Evolutionary Computation Conference*, p. 1651-1656. Morgan Kaufmann, 1999.

[3] Timmis, J., *On parameter adjustment of the immune inspired machine learning algorithm AINE. Technical Report 12-00,.* 2000, Computing Laboratory, University of Kent at Canterbury, Canterbury, U.K.

[4] Tonegawa, S., *Somatic generation of antibody diversity.* Nature, 1983: 302:575-581.

[5] Berek, C. and Milstein, C., *The Dynamic Nature of the Antibody Repertoire.* Immunology Reviews, 1988(105):5-26.

[6] Darnell, J., Lodish, H., and Balitmore, D., *Molecular Cell Biology*. Scientific American Books, 1986.

[7] Hart, E., Ross, P., and Nelson, J., *Producing Robust Schedules via an Artificial Immune System.* Evolutionary Computation, 1998: 6(1):61-81.

[8] Carlier, J., and Pinson, E., *An Algorithm for Solving the Job-Shop Problem.* Management Science, 1989. 35(2):164-176.

[9] Nakano, R., *Conventional Genetic Algorithms for Job-Shop Problems*. In *Fourth International Conference on Genetic Algorithms*, p. 474-479. Morgan Kaufmann, 1991.

[10] Fang, H. L., Ross, P., and Corne, D., *A promising genetic algorithm approach to job-shop scheduling, rescheduling, and open-shop scheduling problems*. In *Proceedings of the Fifth International Conference on Genetic Algorithms*. p. 375-283. Morgan Kaufmann, 1993.

[11] Louis, S., McGraw, G., and Wyckoff, R. O., *Case Based Reasoning Assisted Explanation of Genetic Algorithm Results.* Artificial Intelligence, 1993. 5: 21-37.

[12] Morton, T. and Pentico, D., *Heuristic Scheduling Systems*. John Wiley, 1993.

[13] Holland, J. H., *Adaption in Natural and Artificial Systems.* University of Michigan Press, 1975.

[14] Hightower, R., Perelson, A. S., and Forrest, S., *The Evolution of Emergent Organization in Immune System Gene Libraries*. In *Proceeding of the Sixth International Conference on Genetic Algorithms*. p. 344-350. Morgan Kaufmann., 1995.

[15] Perelson, A. S., Hightower, R., and Forrest, S., *Evolution and Somatic Learning in V-Region Genes.* Research in Immunology, 1996. 147:202-208.

[16] Minar, N., *Suboptimal Solutions in a Simple GA Problem and the Underuse of Genetic Material*. 1994, (unpublished). http://www.santafe.edu/~nelson/gaimmune/gaimmune.html

[17] Oprea, M. and Forrest, S.s *Simulated Evolution of Antibody Libraries Under Pathogen Selection*. In *Proceedings of the 1998 IEEE International Conference on Systems, Man and Cybernetics*. 1998.

[18] Jensen, M., and Hansen, T., *Robust Solutions to Job-Shop Problems*. In *Proceedings of the 1999 Congress on Evolutionary Computation*, p. 1138-1144. IEEE Press, 1999.

[19] Ross, P., and Hart, E., *An Immune System Approach to Scheduling in Changing Environments.* In W. Banzhaf et al. (eds), *Proceedings of the 1999 Genetic and Evolutionary ComputationConference (GECCO)*, p.1559-1565. Morgan Kaufmann, 1999.