

Chapter 5. Static Clonal Selection Algorithm

5.1 Introduction

The problems of negative selection observed in the previous experiments did not originate from the actual nature of negative selection of human immune systems. For the human immune system, the genes of the gene libraries are inherited from ancestors and (through the Baldwin Effect [Baldwin, 1896]), these genes help human immune systems to generate new antibodies to detect antigens that had attempted to attack ancestors' bodies. Returning to our problem, the genes of the initial gene library of the AIS, which will be the genes of pre-detectors, can be selected fields of profiles that describe non-self network traffic patterns. The initial genes might be set by the values of these fields that are observed when a previously known network intrusion is simulated.

This kind of work can be interpreted as providing an automated way of building a misuse detector. When network traffic data is gathered under the two cases where intrusions are simulated and not simulated, the AIS generates detectors containing non-self patterns without overlapping self patterns extracted from these data. This is achieved by the clonal selection algorithm, which lets detectors evolve towards the non-self patterns hidden in the collected non-self data.

This chapter investigates the use of the niching strategy provided by a clonal selection algorithm within the AIS [Kim and Bentley, 2001b]. In addition, in order to solve the scaling problem of an independent negative selection algorithm, the clonal selection algorithm described in this chapter embeds a negative selection operator within it. This new algorithm, called static clonal selection algorithm, is applied to the task of distinguishing self from non-self antigen patterns through a series of experiments and the results of these experiments are analysed in this section.

5.2 Related Work

5.2.1 Clonal Selection of the Human Immune System

Even though new antibodies surviving negative selection are assured to be self-tolerant, their efficacy to detect antigens is unknown when they are released from the bone marrow and the thymus. This is because new antibodies are randomly generated from gene libraries and they are only verified not to be self. They might hold 'non-self' patterns but not 'antigen' patterns. In order to exclude these ineffectual detectors, the human immune system adopts the evolution of antibodies towards the existing 'antigen' patterns [Paul, 1993; Tizard, 1995]. During this evolution process, the human immune system uses its own unique niching strategy to maintain generality and diversity of antibodies as one part of clonal selection process [Forrest *et al.*,

1993]. In a human immune system, this niching process operates only after antibodies are released from the thymus and the bone marrow.

5.2.2 Clonal Selection Algorithms

Forrest *et al.* [1993] presented the niching strategy of their AIS which follows the analogy of the clonal selection in human immune systems. They explored whether it is able to i) detect common patterns of randomly presented antigens and ii) to discern and maintain the diverse antigen population. In their model, they created one population of antibodies and one population of antigens randomly. They used the GA to evolve the antibody population under a constant antigen population. Conforming to the niching strategy of the human immune system, for each generation, their modified GA selects a random sample of arbitrary size from the antibody population and a single random antigen from the antigen population. After each antibody in the sample is matched against a selected antigen, the fitness score of only one antibody showing the highest match score is increased while the fitness scores of the others remain the same.

Using this algorithm, Forrest *et al.* [1993] showed antibodies evolved to be generalists that match most antigens to some extent. Their analysis of this result showed that antibodies evolved towards finding common schemata that are shared among many antigens. Through various experiments, they observed that this algorithm could sustain multiple different antibody patterns, which appear as multiple peaks in a search space, and the similarity among antigens does not affect this capability. Moreover, they compared this niching strategy of the artificial immune system with the fitness sharing algorithm [Smith *et al.*, 1993]. From this comparison, they reported that as the result of the antibody sampling mechanism, the niching strategy of the AIS controls its generality via the antibody sample size. To be more precise, when the sample size decreases, the selective pressures are moved towards generating a population of more general antibodies. Recent work used this algorithm successfully for solving a scheduling problem [Hart *et al.*, 1998; Hart and Ross, 1999].

5.3 Static Clonal Selection Algorithm (StatiCS) Overview

The AIS for network intrusion detection introduced in this chapter adopts the niching strategy of Smith *et al.*'s [1993] AIS. They used a genetic algorithm to construct the AIS. This work modifies this algorithm to be more appropriate for the network intrusion detection problem. This modified algorithm is named the static clonal selection algorithm (StatiCS) since it applies clonal selection on static data collected for misuse detection purposes. Three major modifications were made to StatiCS developed in this work. The first modification is the use of different detector genotype and phenotype representations. Secondly, the fitness and matching functions are altered as the result of detector representation change. Finally, the negative selection stage is embedded in the StatiCS as an operator. The details of these modifications will be described in the following sections. Figure 5.1 provides an overview of the StatiCS.

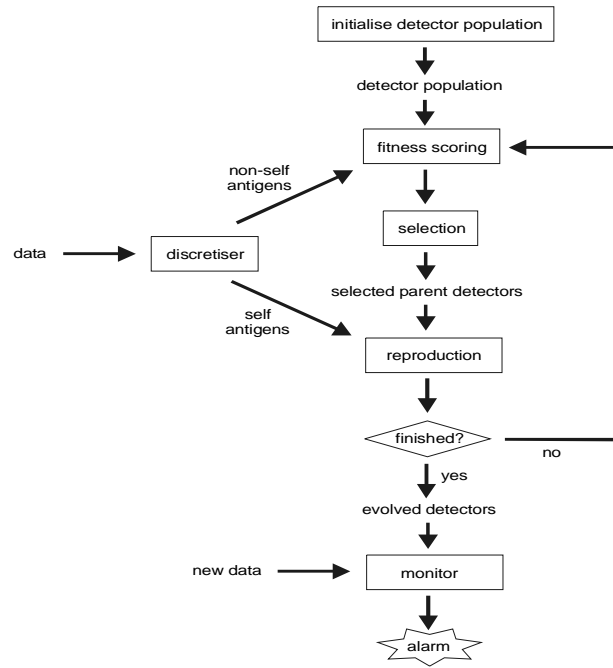


Figure 5.1 An Overview of the StatiCS

5.3.1 Providing Self and Non-Self Antigens

As shown in figure 5.1, when the StatiCS starts, data is fed into the system. In the human immune system, antigens can be divided into two groups: self antigens (our own cells) and non-self antigens (invading pathogens). The clonal selection performed by human immune systems lets antibodies evolve to detect the existing non-self antigens without the detection of any self antigen. The data given to the StatiCS in this work needs to be divided into a self and a non-self set. Since the StatiCS is used for generating an initial detector set, it is assumed that the self or non-self class label is already assigned to each antigen data item. In the case when the data has more than two classes, a single class is predefined as the self and the other classes are regarded as non-self. The self and non-self antigens are then processed by a discretiser before they are passed to the StatiCS.

5.3.2 Discretiser

The antigen data used in this work consists of a number of attributes. These attributes have continuous and discrete values. Specifically, the continuous attribute values often show a wide range of values. Since the detectors generated in the StatiCS employs binary genotypes, a discretisation algorithm is needed. The details of detector genotypes will be discussed in the next section.

There are many discretisation algorithms available and each algorithm has different features [Dougherty *et al.*, 1995]. The StatiCS uses the recursive minimal entropy discretisation algorithm developed by Fayyad and Irani [1993]. This algorithm uses the minimal description

length theory to minimise the entropy between recursively generated intervals. It improved the classification accuracy of c4.5 and Naive-Bayes algorithms on various data sets and it has been known as one of the best general techniques for a supervised discretisation [Witten and Frank, 2000].

Therefore, the continuous value of an attribute for any antigen data will have been clustered into a number of intervals after the discretiser is applied. The range of each interval and the total number of generated intervals are controlled by the discretisation algorithm.

5.3.3 Genotypes and Phenotypes

The StatiCS evolves detectors and these detectors exist as a form of classification rules, which classify non-self from self. A natural expression of classification rules is as a set of disjunctive normal form (DNF) rules. The *if-part* of each rule is a conjunction of one or more conditions to be tested and the *then* side of the rule describes the class label assigned to the rule. In the context of this research, the single detector generated will have a conjunctive rule as its phenotype (Figure 5.2). Therefore, the universal set of non-self patterns that are detected by the detectors is a disjunction of these conjunctive rules.

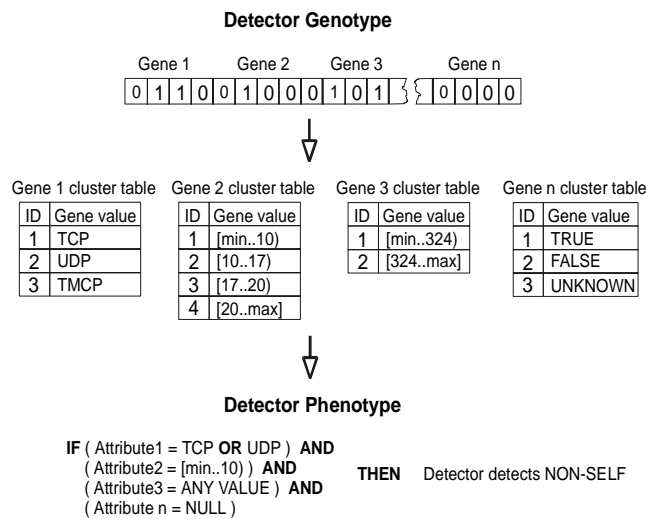


Figure 5.2 Detector Genotype and Phenotype

The StatiCS uses simple binary genotypes in order to encode the conjunctive rule detectors. The StatiCS initialises a detector population by seeding with random genotypes. The detector genotypes consist of a number of genes where each gene represents an attribute of the detector phenotype. The total number of attributes of the given antigen data determines the total number of corresponding genes in the detectors. Each gene is comprised of *nucleotides* and the existing attribute values determine the number of nucleotides. For instance in figure 5.2, in the case of Attribute 1, its valid values are tcp, udp and imcp. Each nucleotide is a binary bit whose value of one represents the inclusion of the corresponding attribute value in the condition part of a classification rule and whose value of zero indicates the omission of the value (see, figure 5.2).

When all bits are zero, the gene is mapped to a value of NULL.¹ This kind of genotype representation allows a single attribute of each detector rule to have more than one value, which are combined by an “OR” operator. In addition, the existing genes of a detector rule are combined by an “AND” operator.

This kind of genotype representation was proposed by De Jong *et al.* [1993] to use a GA for concept learning, which attribute-based classification rules evolve. This encoding scheme has been very popular and its phenotype expression power has been proved to be sufficient enough to handle practical rule evolution problems [De Jong *et al.*, 1993]. On the other hand, this type of genotype and phenotype representation is different from that used by other AIS algorithms for the pattern recognition. The lightweight feature of the detectors generated by other AIS’s [Forrest *et al.*, 1993; Smith *et al.*, 1993; Potter and De Jong, 1998] resulted from the adoption of approximate binding using a matching threshold. The genotypes used in the StatiCS are also expected to have such a lightweight feature because the nucleotides of the genes are connected by OR’s. This phenotype thus allows a detector to detect more than one specific antigen pattern. Therefore, in addition to being lightweight, the detector phenotypes used in this work will have a larger degree of intelligibility. This is because they do not require a numerical threshold whose actual meaning is hard to be understood. When an IDS is designed to send non-self detections to a security officer to confirm and draw a reaction, the intelligibility of detector phenotype is one of the most significant IDS requirements to be satisfied.

5.3.4 The Matching Function

Phenotypes mapped from evolved genotypes are represented in the form of detector patterns. As shown in figure 5.2, an attribute of a detector phenotype is represented by an interval having a lower bound and a higher bound while an attribute of an antigen phenotype is described by one specific value.

Hence, the first step of checking whether a given antigen and a detector match is the comparison of their corresponding attributes. When an antigen attribute value is not within any of the corresponding intervals of a detector phenotype, these two attributes are not matched. For an attribute of nominal type, two genes match when an antigen attribute value is identical to one of the detector phenotype values of its corresponding gene. In order for a given antigen and a detector to match, all the existing genes of the antigen and the detector should match.

There are two special cases: where the first nucleotide of a gene is a one, or where all the nucleotides of a single gene are zeros. The former case will result in the exclusion of this

¹ The first bit of each gene has a special meaning: when it has a value of one, the genotype to phenotype mapping treats the genotype gene as if it is all ones. If it is zero, the remaining bits are used as described. Note that this aspect of the representation was only partially active during tests for vote data, described later, possibly resulting in a slightly degraded TP rate and FP rate. The overall trends were unaffected.

attribute from the condition part of a given classification rule. This is because it is interpreted as any value of this gene is irrelevant to the class decision. In other words, this kind of gene is the generalist gene that match to any value. The latter case is a slightly different. For a network intrusion detection domain, the self and non-self antigens are the observed raw network traffic data. The raw network traffic often produces a null value for a predefined attribute for various reasons such as missing packet. When all the bits of gene have zero values, this detector is matched to the null value of an antigen's corresponding gene.

5.3.5 Fitness Scoring

While the generation of detectors and application of genetic operators are performed at the genotype level, the evaluation of evolved detectors operates at the phenotype level. This is another difference between most work using a negative selection algorithm and clonal selection algorithm [Forrest *et al.*, 1993; 1994; 1997; Hofmeyr 1999; Dasgupta, 1998a]. Such work usually performed this evaluation procedure on a genotype level using a simple r-contiguous bit matching rule. In contrast, here phenotypes mapped from evolved genotypes are represented in a form of detector rules. These detector phenotypes are evaluated by the following fitness scoring procedure. For a non-self antigen set and its corresponding detector set:

1. D detector rules have their fitness values initialised with zeroes.
2. A sample of D detector rules is randomly selected from the generated initial P detector rules.
3. A sample of A non-self antigens are randomly selected from the non-self antigen set.
4. Each detector in the sample is mapped to its phenotype.
5. Each detector phenotype is compared to the selected non-self antigens and the number of matching non-self antigens is counted. This number is defined as a match count for each selected detector.
6. The fitness value of the single detector from the sample that shows the largest match count is increased by the value of the match count. The fitness values of other detectors remain the same. If more than one detector has the largest match count, the fitness value is divided by the number of these tied detectors and their fitness values are increased by the divided fitness value.
7. The processes 2-5 are repeated (for typically three times the number of detectors [Smith *et al.*, 1993]).

As seen in section 5.2.2, this fitness scoring procedure provides the niching strategy for the StatiCS. It controls the generality of each detector according to a detector sample size.

5.3.6 Reproduction and a Negative Selection Operator

After the evaluation of detectors in the detector population, the StatiCS selects parent detectors for the reproduction of detector offspring. The StatiCS uses population overlapping where the

worst $W\%$ detectors are replaced by the best $B\%$ detectors from the newly generated offspring. In addition, a negative selection operator is applied to assure the validity of offspring. This whole reproduction process is described in figure 5.3.

As shown in figure 5.3, the offspring detectors are generated by applying crossover and mutations to two parents randomly selected from the fittest $B\%$ detector rules. The generated offspring are compared to given self antigens. When the offspring matches any self antigen, this offspring is discarded. This kind of invalid offspring can be created because either the parent detectors originally contain some invalid genes or the mutations distort the valid genes of parent detectors. It is not ideal for the StatiCS to ignore the important and valid genetic information of parents unless it is certain that this kind of bad effect originates from the poor genes of parents. Therefore, when an invalid offspring is produced, the StatiCS attempts to generate a new offspring by applying the genetic operators to the same pair of parents until the number of failures to generate valid offspring is less than a predefined negative selection threshold, Nt . When the number of failures to generate valid offspring is more than Nt , the StatiCS selects a new pair of detector parents and produces new offspring. Offspring generation with negative selection continues until it fills up the empty space of the detector population after the worst $W\%$ detectors are deleted.

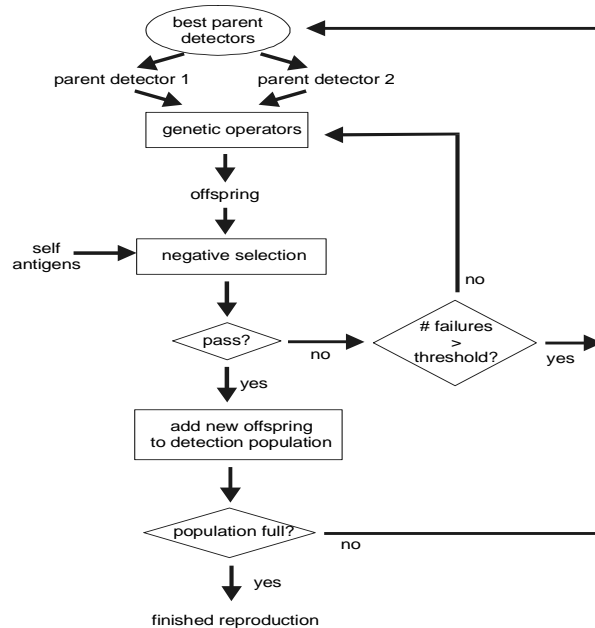


Figure 5.3 Reproduction and Negative Selection

5.3.7 Genetic Operators

The StatiCS applies two genetic operators: crossover and mutation. Since a fixed number of nucleotides represents a genotype, a simple one-point crossover is applied by selecting a random

crossover point between genes or nucleotides. Furthermore, the following five different types of mutations are introduced:

- **Classic mutation:** this mutation is a conventional gene flip mutation. This creates a random mask that has same length bits of a randomly selected gene. Then, it applies an exclusive-OR operator with this mask to the genotype bits of the selected gene.
- **Generalisation mutation:** designed to increase the generality of detectors. It increases the detector generality by causing a new disjunct to be added next to an existing one in the detector phenotype. It selects one bit randomly and if the selected bit has value zero and one of its adjacent bits has value one, then the value of selected bit is changed to the value one. In other words, it is expected to increase the detector generality gradually by adding only an adjacent interval as a new disjunct.
- **Specialisation mutation:** this mutation specialises detectors. This is achieved by dropping a disjunct from detector phenotypes. This is opposite to the generalisation mutation. It selects one bit randomly and if the selected bit has a value one and one of its adjacent bits has a value zero, then the value of selected bit is changed to the value zero. Similarly, this is anticipated to specialise detectors gradually by allowing the dropping of only one adjacent gene of a selected gene.
- **Shift Mutation:** this one shifts all the bits of all the genes to the left or the right direction. The direction to shift is randomly determined. This mutation is designed to give macro variation to whole genotype.
- **Delete Mutation:** this mutation is a stronger version of the generalisation mutation. While the generalisation mutation drops only one disjunction from an attribute and thus brings about the moderate increase of generality, this mutation deletes the whole attribute and results in a rather swift increment of generality. Since the change caused by this mutation is relatively large, the StatiCS controls the effects of mutation by evolution. In order to do it, this mutation flips the first bit of the attribute, changing its corresponding attribute value to 'ANY VALUE' when '1', and back to normal when '0'. This approach is employed to avoid the loss of genetic information which might be useful during future evolution.

These new mutations are mainly introduced to generalise and specialise detectors. This is because the degree of pattern detection of DNF rules is mainly controlled by doing so².

5.4 Experiment Design

5.4.1 Objective

Since the AIS used in all previous work employed a binary detector and simple matching functions (such as Hamming distance or r-contiguous matching), a new investigation was made

in order to understand whether the StatiCS algorithm with the modified detector representation and matching function still allow it to maintain an efficient niching strategy. As introduced in section 5.2.2, the detector sample size controls the generality of detectors generated by the StatiCS. The appropriate mixture of general detectors and specific detectors is critical in order to develop a competent network-based IDS. Detectors should have the maximum level of generality, detecting as many non-self antigen patterns as possible without detecting any self antigen patterns. Furthermore, an ideal detector set should contain detectors showing high specificity that will detect specific antigen patterns found only in a small number of antigens. For these reasons, an ideal detector set should have an appropriate mixture of general detectors and specific detectors. It has been known that the generality of generated detectors is controlled by the detector sample size and the antigen sample size [Forrest *et al.*, 1993; Smith *et al.*, 1993]. With these features of AIS in mind, the experiments were performed to understand how best to choose good detector and antigen sample sizes.

5.4.2 Data and Parameter Setting

This work aims to understand the nature of clonal selection with a negative selection operator. The experiments performed in this chapter did not use real network traffic data sets because such sets are typically vast and are not practically suitable for this type of benchmarking work. Instead three different data sets from the UCI repository for machine learning algorithm benchmark work were used (<ftp://ftp.ics.uci.edu/pub/machine-learning-databases>). While the negative selection algorithm has been tested extensively on various benchmarking data sets [***], the StatiCS type of artificial immune algorithms has not been fully evaluated. This is the main reason for testing StatiCS on the UCI repository data only. The other novel artificial immune algorithms introduced in the later chapters of this thesis will be also tested on the data sets selected from the UCI repository only. Those algorithms will not be evaluated on real network traffic data as the negative selection algorithm has been in chapter 4. For the same reason, evaluating novel algorithms rather than testing matured algorithms, the evaluations of AIS in the remaining chapters of this thesis are limited only to the UCI repository data sets.

The first data set was Wisconsin breast cancer data. It consists of 699 examples with two classes: ‘Malignant’ and ‘Benign’. 241 examples belong to ‘Malignant’ and the rest 458 examples belong to ‘Benign’. ‘Benign’ is defined as a self class and ‘Malignant’ as a non-self class. The detectors generated by the StatiCS detected ‘Malignant’ and any data which was not detected by the detectors was regarded as ‘Benign’. This set had ten continuous attributes and total 16 missing values. The missing values were filled with random values.

The second data set was the ‘vote’ data set. This data set is a collection of voting records and each voting record is classified by one of two parties: ‘Republican’ and ‘Democrat’. It consists

² These mutations are similar to De Jong’s adding and dropping mutations [De Jong *et al.*, 1993].

of 267 democrat and 168 republican examples. Each vote record has 16 voting issues as its attributes and each voting issue has one of three values: yes, no, abstain.

The iris data was used as the final set. It is the most popular data set used in the literature as a pattern recognition test set [Fisher, 1936]. It has total of 150 examples with three classes: 'setosa', 'virginia' and 'versicolour'. Each class has 50 examples and every example has four continuous attributes. Three different data sets are prepared from this original data set by taking one set as a self set and the rest as a non-self set.

A tenfold cross-validation method was employed to prepare a training set for the StatiCS to evolve and a test set to detect previously unseen non-self patterns. The tenfold cross-validation method is known as the most robust method from n fold cross-validations [Witten and Frank, 2000]. A detector population size of 300 was used and best 80% detector offspring were selected to replace the worst 80% detectors from parent detectors, (i.e., 80 was used for both values of B and W). All mutations occurred with a probability of 0.001 per gene. Each experiment was run for a maximum 50 generations unless it satisfied a termination condition. The termination condition was set as a non-self pattern detection rate of 100% and a self pattern detection of 0%. The threshold of the negative selection operator, Nt , was set as 5.

5.5 Experimental Results 1:Non-self and Self Antigen Detection Rates

Two series of experiments were performed by varying the number of detector sample sizes and the number of antigen sample sizes. Other literature suggests that the generality of detectors is controlled by these two factors [Hart and Ross, 1999]. The experiments investigated whether the conclusions of the previous work would be followed in our problem: non-self antigen pattern learning from a collected data set. This section focuses on non-self and self antigen detection rates.

5.5.1 Varying a Detector Sample Size

Table 5.1 and Table 5.2 present the results of the first series of experiments, where the number of antigen samples was fixed and the number of detector samples was varied. The detection rate of the StatiCS was described by a True Positive (TP) rate and a False Positive (FP) rate. TP was the "non-self" detection rate and FP was the rate at which "self" was mistakenly detected by a generated detector set. The desired system should have a high TP rate and a low FP rate. The tables show the means and variances of 10 experiments.

For three data sets, the average TP rates generally showed a good level of accuracy, i.e. more than 93%. For the iris data set the best TP rate reached 100%. There were only a few cases showing less than a 90% TP rate. The average FP rate was consistently lower than 10% for all

cases, but this figure decreased to around 5-6% when D was less than 60 for all three data sets. These good results illustrate the benefits of the StatiCS (compared to those obtained by the negative selection algorithm only). All detector sets were generated in acceptable times: the longest took approximately 35 minutes per set and the shortest took less than a minute.

D	Cancer Data			Vote Data		
	TP (%)	FP (%)	TP-FP (%)	TP (%)	FP (%)	TP-FP (%)
1	93.48 (0.17)	5 (0.26)	88.48 (0.20)	79.43 (0.74)	2.35 (0.09)	77.67 (0.50)
5	94.57 (0.16)	5.83 (0.28)	88.73 (0.36)	88.03 (0.42)	5.29 (0.27)	82.74 (0.84)
10	95.65 (0.12)	5.41 (0.58)	90.23 (0.66)	92.49 (0.40)	3.57 (0.25)	88.93 (0.39)
20	95.43 (0.15)	8.33 (0.73)	87.10 (0.52)	94.02 (0.31)	5.29 (0.27)	88.72 (0.47)
30	95.65 (0.13)	6.25 (0.20)	89.40 (0.27)	93.26 (0.33)	5.92 (0.23)	87.34 (0.62)
60	95.87 (0.13)	9.17 (0.53)	86.70 (0.55)	94.40 (0.28)	5.96 (0.15)	88.45 (0.39)
90				95.16 (0.22)	6.65 (0.26)	88.61 (0.57)
240	96.52 (0.097)	10 (0.548)	86.52 (0.7)	95.55 (0.3)	7.13 (0.3)	88.41 (1.07)

Table 5.1 The mean and variance of true positive rates (TP), false positive rates (FP), and TP-FP rates when an antigen sample size = 1 for various detector sample sizes (D). The mean values are followed by the variances in parentheses.

D	IRIS Setosa			IRIS Versicolor			IRIS Virginia		
	TP (%)	FP (%)	TP-FP (%)	TP (%)	FP (%)	TP-FP (%)	TP (%)	FP (%)	TP-FP (%)
1	100 (0)	0.6 (0.036)	99.4 (0.036)	95 (0.011)	4 (8.889E-03)	91 (0.0289)	95 (0.011)	1 (0.0111)	94 (0.044)
5	100 (0)	0.6 (0.036)	99.4 (0.036)	95 (0.011)	4.8 (0.0196)	90.2 (0.0573)	95.8 (0.0036)	0.012 (1.44E-04)	94.8 (0.019)
10	99.8 (4E-03)	1.2 (0.064)	98.6 (0.063)	95 (0.011)	5 (0.0111)	90 (0.0444)	95.6 (7.11E-03)	1 (0.0111)	94.6 (0.0271)
20	100 (0)	0.6 (0.036)	99.4 (0.036)	95 (0.011)	5 (0.0111)	90 (0.044)	95.6 (7.11E-03)	1 (0.0111)	94.6 (0.027)
30	100 (0)	0 (0)	100 (0)	95 (0.011)	5 (0.0111)	90 (0.044)	95.6 (7.11E-03)	1 (0.0111)	94.6 (0.027)
60	100 (0)	0 (0)	100 (0)	95 (0.011)	5 (0.0111)	90 (0.044)	95.8 (4E-03)	1 (0.0111)	94.8 (0.0196)
240	100 (0)	0.6 (0.036)	99.4 (0.036)	95 (0.011)	4.6 (0.0271)	90.4 (0.0693)	95.4 (9.33E-03)	1 (0.0111)	94.4 (0.0338)

Table 5.2 The mean and variance of TP, FP, TP-FP rates when an antigen sample size = 1 for various detector sample sizes (D). The mean values are followed by the variances in parentheses. IRIS class label in each column indicates the assigned self class.

As table 5.1 explains, the TP rate increased as the detector sample size D increased. From three data sets, the results of the vote data set showed this tendency most clearly. In order to confirm this result, paired sample t-tests were performed on the vote data results. To find the point at which the difference between TP rates becomes statistically significant, t-tests were performed on the pairs of results and each pair was made by taking two adjacent detector sample sizes. The t-test showed that the difference between the TP rates of $D = 1$ and $D = 5$ was statistically significant with 95% confidence. A two-sided t-test of means produced a p-value of 4.3216%. The t-tests of the rest of pairs produced much larger p-values ranging from 14.7285% to 75.385%. In addition, these p-values became larger as the pair was made from larger sample sizes. These results of the t-tests imply that the difference between the average TP rates with varying detector sample sizes converged as the detector sample size increased. Even though the difference of the TP-rate for different sample sizes was very small for the cancer data, the same kind of tendency was observed. However, for the iris data, no results for any D showed any significant difference, see table 5.2.

In addition, the FP rate increased as D increased. The paired sample t-tests were performed on the different pairs which were made in the same way as previous paired sample t-tests. The t-tests showed that the performance difference of the StatiCS between $D = 1$ and $D = 5$ was statistically significant with 94.7% confidence. A two-sided t-test of the means produced a p-value of 5.2177%. Much larger p-values were produced when the t-tests were performed on the rest of pairs, ranging from 35.7729% to 98.7759%. These results also show that the FP rate increased as the detector sample size increased but that it stabilised to a certain point.

5.5.2 Analysis

The observed results were expected. When a detector sample size is one, no niching mechanism can happen. Since there is no chance for a selected detector to compete with other detectors to gain a fitness score, each detector will increase its fitness score by one as long as it matches a given antigen (when $A=1$). Thus, the generalist detector, which detects the largest number of non-self antigens during the fitness scoring procedure, will have the highest fitness score (assuming that each detector is selected with the same probability³). Conversely, more specific detectors will gain much lower fitness scores in the same generation since they will detect much fewer non-self antigens. Thus, the generalist detectors will dominate in a detector population after a certain number of generations.

This kind of phenomenon resulted in rather poor results for the cancer and vote data when $D = 1$. However, the detector sample size did not affect average TP rates for the iris data at all. This is perhaps because the given problem of iris data is relatively easier and thus the minimum sample size is good enough to show a good detection rate. In other words, fairly general detectors can detect all existing non-self antigen patterns in the iris data set.

When the detector sample size is more than one, the selected sample detectors compete with each other. In the tests reported in section 5.5, this led the winner detectors from sampled detector groups to form niches, which match separate peaks of a fitness landscape (see section 5.6 for an analysis of the formation of niches). In the extreme case, when the detector sample size is the largest possible (the detector population size), every detector participates in a competition to detect a given antigen. This gives a chance for very specific detectors to increase their fitness scores because some specific non-self antigen patterns can only be detected by these kinds of detectors. Therefore, these specific detectors will have fitness scores that are large enough not to be excluded from the parent population through selection. In other words, both the general detectors and specific detectors have fair chances to win and thus they both will remain in the final detector population.

³ This is a reasonable assumption because detector sampling during fitness scoring is repeated a sufficient number of times to ensure its validity (typically three times the number of detectors).

However, when a detector sample size is the largest possible, it can cause an overfitting problem. The specific non-self antigen pattern may not be representative of the data as a whole. So a detector evolved to match this exceptional antigen pattern might not truly distinguish between “self” and “non-self”, resulting in higher false-positive rates. This overfitting problem is clearly observed from the experimental results shown in section 5.5. For both data, cancer data and vote data, the FP rate increases as the detector sample size increases, see table 5.1 and table 5.2.

5.5.3 Varying a Antigen Sample Size

A	Cancer Data			Vote Data		
	TP (%)	FP (%)	TP-FP (%)	TP (%)	FP (%)	TP-FP (%)
1	95.65 (0.12)	5.42 (0.58)	90.23 (0.66)	92.49 (0.40)	3.57 (0.25)	88.93 (0.39)
5	94.35 (0.18)	3.75 (0.40)	90.6 (0.31)	92.14 (0.44)	3.54 (0.07)	88.59 (0.42)
10	95 (0.16)	5.42 (0.39)	89.58 (0.35)	89.56 (0.39)	2.94 (0.17)	86.62 (0.75)
MAX	93.91 (0.24)	5.42 (0.31)	88.5 (0.24)	85.47 (1.63)	3.57 (0.17)	81.90 (2.17)

Table 5.3 The mean and variance of TP, FP, TP-FP rates when a detector sample size = 10 for various antigen sample sizes (A). The mean values are followed by the variances in parentheses.

Next the results were compared when the detector sample size was fixed but the antigen sample size changed. The last series of experiments were performed with $D = 10$ and various antigen sample sizes⁴. As seen in table 5.3, no significant difference between TP’s and FP’s was evident, except for the case where the antigen sample size was the maximum.

5.5.4 Analysis

These results are also readily explainable. When the antigen sample size is small, even a potentially general detector does not have enough opportunity to detect a large number of antigens and thus both a general detector and a specific detector will be compared only for whether they can detect a given small number of antigens. Thus, the difference of fitness scores is not large. However, as the antigen sample size increases, the general detector starts to have enough chances to beat the specific detector by detecting a larger number of antigens. Thus, the general detectors have more chances to be selected as the parents for the next generation. So larger antigen sample sizes can also cause domination of general detectors during evolution.

5.5.5 Ideal Detector Sample Size and Antigen Sample Size

Since an ideal IDS should show a high TP rate and low FP rate, TP-FP rates were analysed to take into account these two rates together. As shown in table 5.1, for cancer data, these rates did not show significant differences for any case. For the vote data, it stabilised after a detector sample size reached 10. These results advise that the detector sample size does not have to be the largest one to get the most ideal result. Instead, a detector sample size that was used in the experiments is not too small but is large enough to gain the good TP-FP rate. To be more

precise, experimental results suggest that the detector sample size should be set as the largest size which is affordable by given system resources. As future work, an adaptive sample size determined through evolution can also be investigated. This approach will be beneficial when it is considered that the larger detector sample size can cause an overfitting problem.

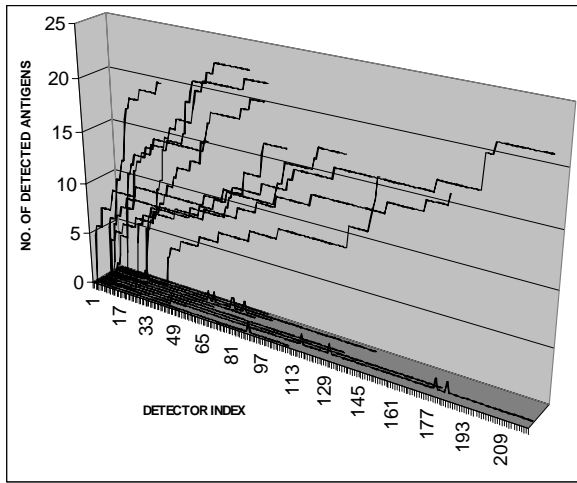
As long as the detector sample size is properly set, the antigen sample size is not critical. Since our experiment results show that the generality of detectors can be controlled by the detector sample size, the smallest antigen sample size ($A = 1$) is recommended. This is because the minimum antigen size saves computation time.

5.6 Experimental Results 2: The Degree of Generality in a Detector Population

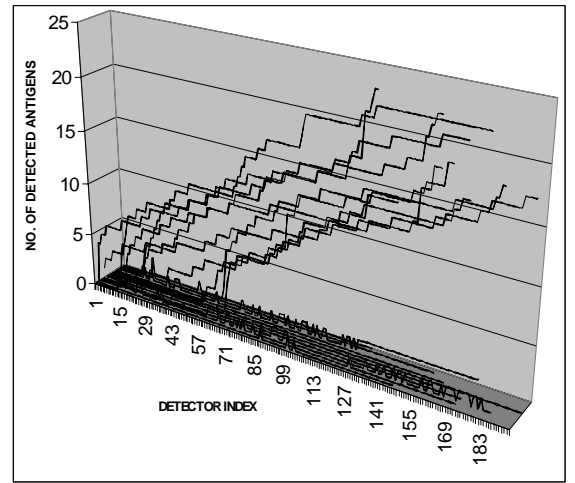
As emphasized in the previous section, a smaller detector sample size causes the evolution of the StatiCS to be dominated by more general detectors and a larger detector sample size allows more specific detectors to remain in a final detector population. To confirm this feature, the degree of generality of detectors is analysed. This analysis shows the distribution of general and specific detectors in a final detector population. The degree of generality for each detector is defined by the number of different antigens detected by each detector. For instance, if any detector detects a very large number of different antigens, this detector shows a large degree of generality and if it detects a small number of unique antigens, then it indicates a low degree of generality.

Figure 5.4 shows various degrees of detector generality when the detector sample size changes and the antigen sample size is fixed at one, when the vote data set was used for generating detectors. The X-axes of these graphs indicate detector indexes. For instance, from the first graph showing the degrees of generality for the detectors in a final population with $D = 1$ and $A = 1$, the X-axis ranges from 1 to 217. This means that the number of different detectors in the final population is 217. The Y-axes show the numbers of antigens detected by each detector. These antigens are selected from only a test antigen set. The graphs were drawn as follows. The detectors were sorted according to the ascending order of detected antigen numbers. According to this order, the unique detector indexes were allocated. Then, the lines were drawn by pairing a given detector index to the corresponding number of detected antigens. Each line describes the self antigen or non-self antigen detection results of a single experiment from ten-fold cross validation. Each graph always shows twenty different lines. These twenty lines are divided into

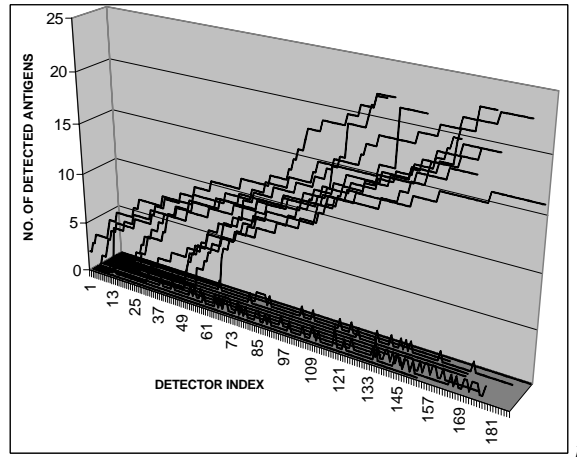
⁴ From the previous experiments performed in section 5.5 the problem of detecting self and non self IRIS data is easier, and thus experimental results did not show noticeable differences depending on the detectors' sample size. For this reason, the experiments to investigate the effects of various antigen sample sizes use only cancer and vote data.



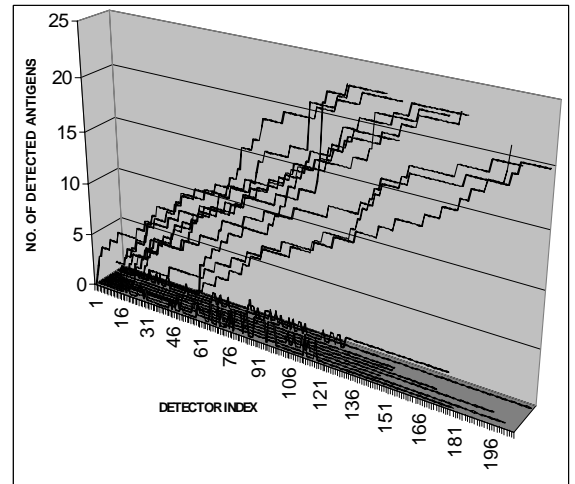
$D = 1, A = 1$



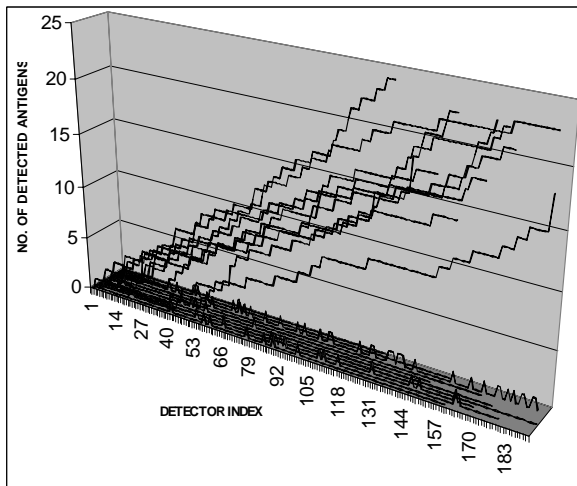
$D = 5, A = 1$



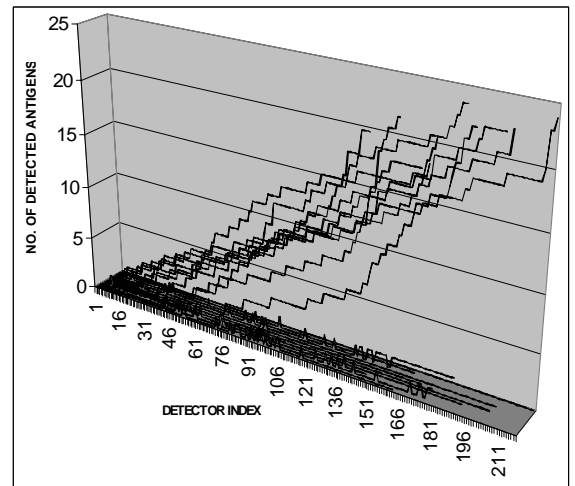
$D = 10, A = 1$



$D = 20, A = 1$



$D = 30, A = 1$



$D = 90, A = 1$

Figure 5.4 Degrees of detector generality when the detector sample size changes and the antigen sample size is fixed at one.

two groups. The first ten lines that remain near the bottom of each graph describe the number of detected self antigens, and the remaining ten lines present the number of detected non-self antigens. The lines presenting the number of detected self antigens are distributed evenly across a Z-axis for clearer viewing of lines that would otherwise overlap in a 2-D graph.

These graphs are interpreted as follows. Since the lines were drawn by connecting points that indicate specific numbers of detected antigens, these specific numbers form the flat parts of each line. Thus, the width of these flat parts of each line represent how many different detectors show the same degree of generality. Because the detectors were already sorted according to the number of detected antigens, if these flat parts are wider, it implies that the StatiCS generates a larger number of detectors that have the same degree of generality.

From these graphs, it can be seen that the proportion⁵ of detectors showing the same generality decreases as the detector sample size increases. From the first graph, which shows this trend when $D=1$ and $A=1$, most of lines have a large proportion of detectors showing the same generality, which is shown by having a larger number of wider flat parts and a smaller number of narrower flat parts. From the next graph, when $D=5$ and $A=1$, this large proportion of detectors showing the same generality has already started to decrease. In contrast, when $D=90$ and $A=1$, the proportion of narrower flat parts (indicating a smaller number of detectors showing the same generality) is clearly much higher than in any of the other cases. In addition, the change of these proportions from the minimum detector sample size ($D=1$) to the next smallest sample size ($D=5$) is a lot larger than the change from the second largest detector sample size ($D=30$) to the largest one ($D=90$). These observations corroborate our previous understanding that some degree of niching occurs as long as more than one detector sample is used for fitness scoring (see section 5.5.2). To be more precise, there is no niching mechanism when $D=1$, resulting in very general detectors, but some niching when $D=5$, resulting in a big change to more specific detectors. Thus this case shows the most dramatically different figure compared to the rest cases.

Furthermore, let us consider the charts showing wider flat parts of lines and where these parts of the lines appear in the charts. As stated before, the wider flat parts of graphs describe a larger proportion of detectors that show the same generality. Thus if these wider flat segments occur at the top of graphs, this suggests that a larger number of very general detectors have been generated. In contrast, if these wider flat parts of lines appear at the bottom of graphs, it means that a larger numbers of specific detectors have been produced. However, no particular trend about the location of wider flat parts of lines is apparent. Before the detector sample size

⁵ Each experiment produced a different number of detectors in the final detector population and thus X axes ranges of lines are quite different. Therefore the proportion of detectors showing the same generality is defined as “total

reaches the point when the StatiCS starts to generate a very diverse generality of detectors, which is about $D \leq 30$, the wider flat parts appear at both the bottom and the top of graphs. Hence, a more precise conclusion about the effect of different detector sample size can be drawn from these results: as the detector sample size increases, the StatiCS creates more diverse types of detectors in terms of their generality. This again implies that a more precise mixture of general and specific detectors follows fitness landscape more closely as the detector sample size increases.

Before this observation, what was understood from the previous analyses of TP, FP and TP-FP rates was that a smaller sample size causes the domination of very general detectors and a larger sample size results in the remains of specific detectors in a final detector population. However, it was not clear that whether this understanding explained the exact distribution of general detectors and specific detectors when a detector sample size changes. However, the new observation show in figure 5.4 suggests that a larger detector sample size leads the evolution of StatiCS to maintain various niches and that each niche also shows diverse degree of generality. In other words, various detectors having a diverse degree of generality are distributed very evenly in a final detector population when a detector sample size becomes large.

This explanation provides more evidence of an overfitting problem which occurs when the detector sample size increases. Since a large detector sample size encourages the StatiCS to produce a more precise mixture of general and specific detectors, the final detector population follows the fitness landscape more closely for larger values of D . Therefore, the bottom ten lines showing the number of detected self antigens get more spiky as the value of D increases. This is because it detects more self antigens, see figure 5.4.

5.7 Experimental Results 3: The Performance of the Negative Selection Operator

The experiments on the StatiCS, combined with the earlier experiments on the traditional negative selection algorithm confirm that the role of negative selection should be to support clonal selection. From the experiments performed in chapter 4, it was observed that the negative selection algorithm suffers from severe scaling problems when applied to realistic network traffic data. From the experiments performed in this chapter, it was observed that the negative selection operator played an important role which helped to reduce the FP rate for the StatiCS. When evolution terminated at the maximum generation and the detectors were tested on a training data set, no case showed any mistake, i.e., FP was always 0% on the training data set. For the test set, the observed FP rate was up to about 10%.

number of different detectors showing the same generality / total number of different detectors in the final detector population”.

As discussed before, the FP rate is mainly controlled by a detector sample size and an antigen sample size. Therefore, the rather higher FP rates resulted not because of inappropriate behaviours of the negative selection operator but because of the improper choices of detector sample sizes and antigen samples sizes. However, it have not been investigated how the threshold size of negative selection operator will affect the TP and FP rate. Too small a threshold size might lead to prevent the generation of some general detectors because it will eliminate detectors matching very small number of self antigens. However, these self antigens can be noise. Similarly, too large a threshold size can make the StatiCS to generate the detectors which are so general that they detect too many self antigens. Thus, the effect of negative selection threshold size should be investigated as the future work.

5.8 Summary

This chapter was devoted to the investigation of the use of a clonal selection algorithm with a negative selection operator, called StatiCS. StatiCS was specially developed for the purpose of building a misuse detector in an efficient way. In order to adapt the available clonal selection algorithm for a network intrusion detection application, three major modifications were made to the StatiCS: i) new genotype and phenotype representations, ii) new matching and fitness score functions, and iii) introduction of a negative selection operator.

Among these modifications, the new phenotype representation (representing a conjunctive rule) removes the matching threshold parameter by allowing an “OR” operator in a genotype-phenotype mapping. The matching threshold parameter has been known to greatly affect the detection rate of the original negative selection algorithm, but the choice of parameter values has been made arbitrarily. In contrast, the modified phenotype representation can dynamically alter the matching scope of a given detector since it embeds an “OR” operator in the phenotype. Thus, the modified phenotype representation no longer requires the arbitrary choice of a parameter value that significantly affects the detection rate. In addition, this phenotype allows a detector to detect more than one specific antigen pattern, and thus it retains the lightweight feature originated from approximate binding. Furthermore, the detector phenotypes used in this work will have a larger degree of intelligibility. This is because they do not require a numerical threshold whose actual meaning is hard to ascertain. When an IDS is designed to send non-self detections to a security officer to confirm and draw a reaction, the intelligibility of detector phenotypes is one of the most important requirements of the IDS.

The experimental results from the StatiCS showed a good non-self antigen detection rate with a low self antigen detection rate. In contrast with the results obtained for the negative selection algorithm, these results were gained in acceptable times. These results are valid for the environment used for experiments performed in this chapter. In order to advocate the same

results in the real network environment, further tests should be conducted in future. Two series of experiments were performed by varying the detector sample size and the antigen sample size. These experiments were executed in order to investigate the effect of detector and antigen sample sizes on performance. The results of these experiments showed a good non-self antigen detection rate with a relatively large detector sample size. These results also suggested that a more precise mixture of general and specific detectors follows the fitness landscape more closely as the detector sample size increases.

This work has shown that clonal selection provides competent non self antigen detection and the most appropriate use of negative selection in the AIS is as a filter for invalid detectors, not the generation of competent detectors.