Negative Selection Algorithm for DNA Sequence Classification

Dong-Wook Lee and Kwee-Bo Sim
School of Electrical and Electronic Engineering, Chung-Ang University
221, Heukseok-dong, Dongjak-gu, Seoul 156-756, Korea
Email: dwlee@wm.cau.ac.kr, kbsim@cau.ac.kr

***Abstract*— We propose a pattern classification algorithm using self-nonself discrimination principle of immune cells and apply it to DNA pattern classification problem. Pattern classification problem in bioinformatics is very important and frequent one. In this paper, we propose a classification algorithm based on the negative selection of the immune system to classify DNA patterns. The negative selection is the process to determine an antigenic receptor that recognize antigens, nonself cells. The immune cells use this antigen receptor to judge whether a self or not. If one composes n groups of antigenic receptor for n different patterns, these receptor groups can classify into n patterns. We propose a pattern classification algorithm based on the negative selection in nucleotide base level and amino acid level. Also to show the validity of our algorithm, experimental results of RNA group classification are presented.**

I. Introduction

By the growth of the molecular biology and the success of the genome project, we can obtain a DNA sequence of human and other living things. However, a DNA sequence provide no immediate information as "Which parts are genes?" "How and when is revealed a gene?" at all. Also, about 10 percent of human genome has genetic information that synthesizes proteins, and this gene area is distributed in genome. Therefore, to utilize the DNA sequence obtained by genome project, post-genome research that includes interdisciplinary field such as biology, computer science, mathematics, statistics, and information theory has been started in recent year. A representative field of this research is bioinformatics [1].

In this paper, we propose a DNA pattern classification algorithm based on the negative selection of Biological Immune System (BIS). BIS is complex and sophisticated system to recognize and eliminate antigens from outside [2], [3]. BIS generates various antibodies to recognize foreign antigens. Antibody producing immune cell is B cell. B cells are generated through the negative selection not to recognize self as nonself. B cells which get through the negative selection have the classification ability of self and nonself.

Another representative immune cell is T cell. T cells have two type of receptor. One is antigenic receptor that is used for recognizing antigen, and the other is MHC receptor that is used for recognizing self molecule. Each receptor of T cells is obtained through the negative selection and the positive selection. These selection mechanisms of BIS have been modeled to various engineering applications. Forrest *et al.* [4], [5], [6] proposed the negative selection algorithm for applying it to anomaly detection in computer system. Kim and Bently [7] utilize the negative selection for network intrusion detection. Sim and Lee [8] developed self-nonself recognition algorithm based on positive and negative selection and Esponda *et al.* [9] presented a formal framework for positive and negative detection schemes.

In our research, we develop a pattern classification algorithm using self-nonself discrimination principle of immune cells and apply it to DNA pattern classification problem. Pattern classification problem in bioinformatics is very important and frequent problem. For example of these problems are identification of coding regions in DNA sequences, classification of protein group, classification of RNA group, analysis of microarray of DNA chips, structure prediction of protein and RNA, analysis of gene revelation, and so on [1]. To solve these problems, machine learning approach such as neural networks, evolutionary computation, probabilistic graph model are applied in recent researches [10], [11], [12]. In this paper, we propose a pattern classification algorithm based on negative selection in nucleotide base level and amino acid level. We also propose a detector generating method based on Genetic Algorithms (GA). Finally, to show the validity of our algorithm, experimental results of RNA group classification are presented.

II. DNA Structure

Natural living things have their own DNA (deoxyribonucleic acid) sequences in cell [2]. DNA is a genetic code that emerges to the characteristics of individual. Biological DNA consists of nucleotides which have Adenine (A), Thymine (T: Uracil (U) in RNA), Guanine (G), and Cytosine (C). A messenger RNA (ribonucleic acid) is first synthesized from DNA. Three successive bases called codon is allocated sequentially in the mRNA. These codons are the codes for amino acids. Sixty-four kinds of codon correspond to 20 kinds of amino acid (Table I). The allocations of amino acid make proteins and proteins make up cells. Translation of mRNA starts on AUG, and comes to an end on UGA (or UAA, UAG). So, only the nucleotide bases that exist between start codon (AUG) and stop codon (UGA, UAA, UAG) are translated into a protein.

Fig. 1 shows an example of DNA translation. One codon codes one amino acid and the location of stat codon decides the pair of 3 nucleotide bases. So, a DNA can be translated

TABLE I
GENETIC CODE.

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UGG		UAG		UGG	Trp	G
C	CUU		CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CGG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	AGG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GGG		GAG		GGG		G

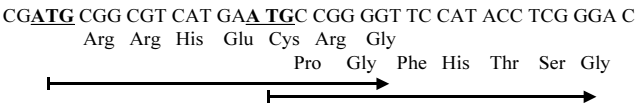

Fig. 1. An example of DNA translation.

by 3 methods according to the location of start codon. These 3 types of translation method is called reading frame.

III. NEGATIVE SELECTION ALGORITHM BASED ON BIOLOGICAL IMMUNE SYSTEM

A. Biological Immune System

Biological immune system (BIS), the protection system of living creature, is so complex and sophisticated system to protect cells and organs from various external organisms or proteins that are named as antigens, such as bacteria, pathogens, and viruses. The basic elements of immune system are two types of lymphocytes, B cells (B lymphocytes) and T cells (T lymphocytes). B cells take part in humoral response that secretes antibodies, and T cells take part in cell mediated immunity that stimulates or suppresses cells concerned with immune response and kills infected self cells. The protein that represents each characteristic exists in the individual. It is called MHC (Major Histocompatibility Complex) molecule. The part to recognize this MHC molecule is located in an immune cell's body. The immune cell uses this protein to judge whether a self or not. Also, the immune cell such as B cell or T cell has detector that recognizes specific antigen. This is called antigenic receptor [2], [3].

In BIS, an immune cell being core of the immune response relies on two elements to eliminate antigens that intrude a living body. One is cooperation and communication between cells. The other is the ability to recognize an antigen and discriminate between self and nonself. A representative immune cell is cytotoxic T cell that has both antigenic receptor to recognize antigens and MHC receptor to recognize the MHC molecules (MHC proteins) identifying a self-cell. Cytotoxic T cells are produced through positive and negative selection. If a T cell receptor doesn't operate properly in the immune system, it recognizes a self-cell as an antigen and attacks it. Therefore, when T cells are produced initially, it is examined the proper operation of MHC receptor and antigenic receptor. These processes are the positive and the negative selection. They judge whether two receptors are operated properly or not.

The positive selection is a way to examine MHC recognition function of each immature immune cell. Because only immature immune cells which can recognize MHC molecules correctly in the self-cell can be used for immune system. Mature immune cells consist of only cells that are selected positively among immature immune cells that are matched with MHC molecules. At this time, the immune system can be maintained by elimination of the unmatched cells, because the unmatched immature immune cells can't recognize a self-cell.

The negative selection is a way to exclude immature immune cell recognizing a self-cell as an antigen. If an antigenic receptor recognizes MHC molecule as antigen, the antigenic receptor takes all of self-cells as antigens. When an immature immune cell combines to MHC molecule, only a cell that has antigenic receptor not to recognize MHC molecule as antigen is selected to a mature immune cell. If the immature immune cell to select it positively recognizes MHC molecule as antigen, it is eliminated.

The immature immune cells form a proper immune response in a living thing after completing these two selections. The generating process of immune cells is shown in Fig. 2.

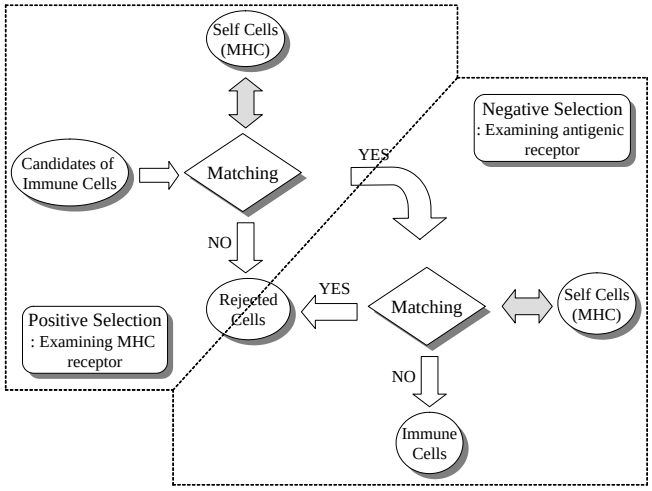

Fig. 2. Generating process of T cells.

B. Negative Selection Algorithm

The anomaly detection algorithm based on the negative selection is one of self-nonself discrimination algorithm that

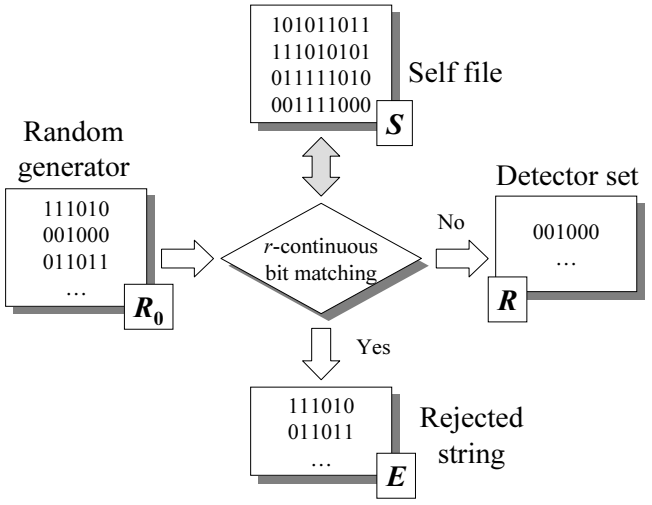

Fig. 3. Generation of anomaly detectors using negative selection.

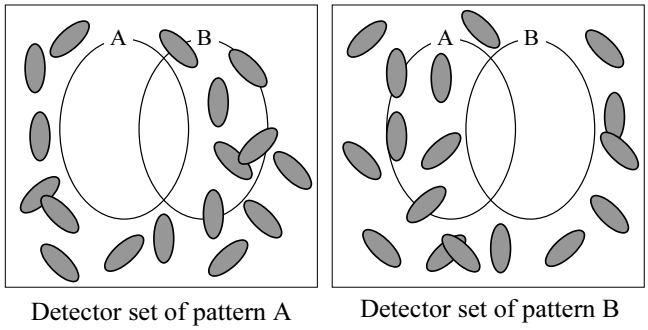

Fig. 4. Pattern classification using detector set.

was proposed by Forrest *et al.* [4], [5], [6]. They composed the set of detectors that don't recognize self-space. Composed detector set is used for nonself recognition. This algorithm is divided into two parts. One is the part to compose anomaly detectors by the negative selection. The other is the part to check the modification of self by composed detector set.

Fig. 3 represents the process to produce anomaly detectors by the negative selection. Anomaly detector is compromised using strings that are not matched to self-space. First of all, define a self-space S that should be protected. The next, make a random string which length is l. Let the set of random string is D_0. After r-contiguous matching between each string in D_0 and all strings in S, we can compose detector set D, which is the collection of unmatched strings in D_0, where $r < l$. At this time, matched string is rejected.

The perfect matching between the two strings having a same length means that all the same symbol of each cell is located in each position of string. Because this matching is difficult to find string that isn't matched as self-string gets larger, a partial matching rule is used. The matching rule which the anomaly detection algorithm uses is an r-contiguous matching rule. If two strings have same r-contiguous cells, they are regarded as being matched.

We can recognize self and nonself using the anomaly detectors that were made through above process. This algorithm has a merit that it can recognize various antigen, modification, by preparing sufficient anomaly detectors.

IV. Pattern Classification Algorithm based on Negative Selection

A. Classification in Nucleotide Base Level

In this section, we explain our pattern classification algorithm based on the negative selection. The negative selection is the way to determine an antigenic receptor that doesn't recognize self. Immune cells that have antigenic receptor can classify self from nonself. Likewise, immune cells, which are generated by the negative selection of a specific self, have ability to recognize the specific self. Therefore these immune cells can be used as detectors to recognize self. If we regard self as a pattern and extend this analogy to n patterns, we can obtain n detector sets that determine each n pattern. Fig. 4 is the diagram that shows pattern A, pattern B, and their detectors.

DNA consists of 4 nucleotide base. So, 4-base system is useful to treat DNA data. Instead of r-contiguous matching rule of the anomaly detection, we use hamming distance between two DNA strings. If a character of specific locus of two strings are same then hamming distance increase 1, otherwise it doesn't change. Selection principle is to select a detector that has big hamming distance between the detector and self strings. We introduce (1) and matching threshold r for selection.

$$H(d, s_i) > r \tag{1}$$

where $H()$ is hamming distance, d is a randomly generated detector, s_i is the i-th sub-string of S, and S is the self string of a pattern. The length of string d and s_i are l.

Fig. 5 shows the composition method of detector sets for each pattern using the negative selection. By above process, we can obtain n detector sets for n patterns. We can classify an input pattern using these detector sets.

The time taken to generate the detectors is measured by the number of candidate detectors that have to be examined before producing the required number of competent detectors. It was observed that the number of candidate detectors increases exponentailly with the size of the self-set and the length of detector [13]. Original negative selection algorithm by Forrest [4] used the exhaustive detector generating algorithm. This algorithm attempts to construct a set of competent detectors as shown in Fig. 3. This algorithm take long generation time and dose not check for redundent detectors. To improve these limitations, some variations of detector generating algorithm were developed: linear [6], greedy [6], and binary template [14]. These algorithm minimize the time taken to generate detectors, but it still take long time as the increase the size of self data and the lengrh of detector. Besides, GA is the

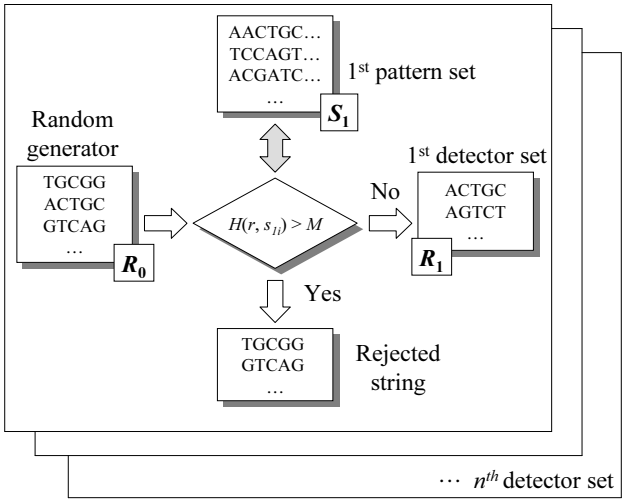

Fig. 5. Composition method of detector sets for each pattern.

powerful tool to search the solution space using stochastic process. So, we propose GA based detector generating algorithm as follows. Our algorithm utilize fitness sharing function [15] for the diversity of detectors.

[Detector Generation Algorithm based on GA]

1) Initialize detectors at random.
2) Evaluate the fitness of detectors based on matching function and sharing function using (2).

$$f_i = \frac{m_i}{\sum_{i=1}^{N} sh(h_{ij})} \tag{2}$$

where m_i is the matching function (hamming distance or contiguous matching) value of ith detector and a self pattern, N is the size of population, $sh(h_{ij})$ is sharing function which is defined as (3).

$$sh(h_{ij}) = \begin{cases} 1 - \dfrac{h_{ij}}{\sigma_s}, & \text{for } 0 \le d_{ij} < \sigma_s \\ 0, & \text{for } d_{ij} \ge \sigma_s \end{cases} \tag{3}$$

where h_{ij} is the hamming distance between individual i and j, σ_s is the parameter which represents sharing radius.

3) If best n individuals out of N are satisfied with (1) and their fitness are all above the predefined value, then goto 5).
4) For all individuals, crossover and mutation operation are performed according to crossover and mutation probability.
5) A roulette wheel selection are adopted to compose the next population.
6) Compose the set of detectors for recognizing the self pattern.

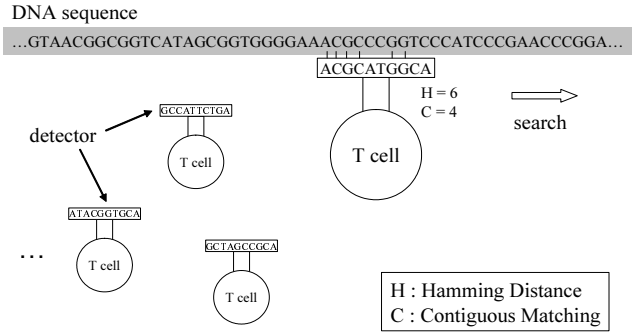

Fig. 6. DNA sequence recognition using (T cell) detectors.

When a new pattern is entered, it is tested by the composed detector set. Detailed pattern recognition method is as follow. If maximum hamming distance between all detectors of specific pattern and an input pattern is below the matching threshold r, then the input pattern is regarded as the specific pattern. If not, then the input pattern is not in the specific pattern. Equation (4) is the decision function.

$$max_{k,i}[H(d_k, s_i)] \le r \tag{4}$$

where k is the index of the detector.

Fig. 6 shows the recognition process of DNA sequence using composed detectors.

B. Classification in Amino Acid Level

If we are to solve the classification of protein group or structure prediction of protein, it is more useful to use amino acid unit instead of 4-base unit. So, It is needed to translate DNA sequence to amino acid sequence before applying it to classification. Because the number of amino acid is 20, we use 20-ary system. Table II shows 20 amino acid codes. The pattern classification method is same as in nucleotide base level.

TABLE II
AMINO ACID CODE.

No.	1	2	3	4	5	6	7	8	9	10
code	A	R	N	D	C	Q	E	G	H	I
A. A.	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
No.	11	12	13	14	15	16	17	18	19	20
code	L	K	M	F	P	S	T	W	Y	V
A. A.	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

V. EXPERIMENTAL RESULT

To verify the effectiveness of pattern classification, the proposed method was tested on a data set of rRNA sequences. rRNA is ribosomal RNA that is an organ to synthesize proteins in cell. The pattern of rRNA sequence is different according to the species. Bacteria as procaryote and fungi as eucaryota have

their own rRNA patterns. We use actinobacteria (procaryote) data and basidimycota (eucaryta) data for experiment. Experimental data are obtained from comparative RNA web site [16]. rRNA is classified into three types by length. These are 5S, 16S, 23S rRNA respectively. In experiment, we use 5S rRNA which has about 120 nucleotide bases. We obtained 48 actinobacteria rRNA sequences and 36 basidiomycota rRNA from [16]. So we set the number of training patterns as 24 and that of test patterns as 24 and 12 respectively. Also, to calculate the FNR (False Negative Rate), we randomly generate 100 rRNA sequences which are in neither actinobacteria nor basidiomtcota. To obtain the general performance of classification, we experiment 10 times and calculate average performance. The detector set of each pattern is regenerated at every time. Also we compare our hamming distance based method with r-contiguous matching method.

Table III shows the parameters for pattern classification. We use 10 detectors and set the length of detector as 10. Matching threshold (r) is 5 for hamming distance based method and 3 for r-contiguous matching method. In this experiment, one datum has about 120 nucleotide bases. So one detector matches 110 times, because a detector is shifted 110 times for matching with an input pattern (see Fig. 6). In recognition phase, we allowed the excess of matching threshold (tolerance: e) as 3 (or 5) out of 110. In detector generation algorithm, the parameters are set as follows: number of population (N) is 50; crossover probability is 0.8; mutation probability is 0.1. In every experiment, finally obtained generation was not exceed 50.

Matching threshold is selected to the minimum value as possible, because a detector covers more broad area of a pattern space as a matching threshold is small. The parameters of Table 3 are determined heuristically. In experiment, the performance of pattern recognition depends on the parameter setting. Therefore to improve the performance of this algorithm, it is needed to optimize the value of the number of detector and the length of detector.

Tables IV and V represent the results of experiment. When the tolerance is 3, we obtain the recognition rate as 93.4% and FNR as 0.0% in the recognition of actinobacteria pattern (Ac. in table), and the recognition rate as 82.7% and FNR as 0.8% in the recognition of basidiomycota pattern (Ba. in table). When the tolerance is 5, we obtain the recognition rate as 97.4% and FNR as 0.8% in the recognition of Ac. pattern, and the recognition rate as 91.8% and FNR as 3.1% in the recognition of Ba. pattern. As the tolerance is increase, the recognition rate is also increase, but the FNR is decrease.

TABLE III
EXPERIMENTAL PARAMETERS.

matching method	# of detector	detector length (l)	threshold (r)	tolerance (e)
Hamming D.	10	10	5	3/5
Contig. M.	10	10	3	3/5

TABLE IV
EXPERIMENTAL RESULT OF DNA PATTERN RECOGNITION WHERE THE TOLERANCE (e) IS 3.

matching method	pattern name	# train patterns	# test patterns	r. rate[a] (%)	FNR (%)
Hamming Distance	Ac.[b]	24	24	93.4	0.0
	Ba.[c]	24	12	82.7	0.8
Contiguous Matching	Ac.	24	24	90.0	1.0
	Ba.	24	12	75.5	1.1

[a]recognition rate, [b]Actinobacteria, [c]Basidiomycota

TABLE V
EXPERIMENTAL RESULT OF DNA PATTERN RECOGNITION WHERE THE TOLERANCE (e) IS 5.

matching method	pattern name	# train patterns	# test patterns	r. rate (%)	FNR (%)
Hamming Distance	Ac.	24	24	97.4	0.8
	Ba.	24	12	91.8	3.1
Contiguous Matching	Ac.	24	24	92.2	4.6
	Ba.	24	12	88.2	3.0

In order to examine the classification ability of the proposed algorithm. We acquire the mutual recognition rate between two patterns. Table VI shows the results of mutual recognition rate by hamming distance based method and r-contiguous matching based method. In hamming distance based method, the detector set trained by Ac. patterns does not recognize 1 datum out of 360 data (0.28%). In this case, the detector set recognizes Ba. pattern as Ac. pattern. Also, the detector set trained by Ba. patterns don't recognize 34 data out of 480 data (7.08%). In this case, the detector set recognize Ba. pattern as Ac. pattern. Therefore, in the classification for these two patterns, we can expect the classification rate to be about 92.92% at the minimum.

In these experiments, the performance of hamming distance based method is superior than r-contiguous matching based method.

VI. CONCLUSION

In this paper we propose the pattern classification algorithm based on the negative selection of BIS. The negative selection is the method to generate immune cells that can discriminate self and antigen. In our research, we developed

TABLE VI
MUTUAL RECOGNITION RATE OF ACTINOBACTERIA AND BASIDIOMYCOTA PATTERN.

matching method	training pattern	test pattern	mutual r. rate (#/total #)
Hamming Distance	Ac.	Ba.	0.28 (1/360)
	Ba.	Ac.	7.08 (34/480)
Contiguous Matching	Ac.	Ba.	0.56 (2/360)
	Ba.	Ac.	11.67 (56/480)

pattern classification method by introducing the self-nonself discrimination method. This is implemented by composing n detector sets for n patterns. Conventional pattern classification method needs the agreement of pattern size and the feature extraction process. However proposed method doesn't need the feature extraction and the agreement of pattern size. It is also effective when a pattern size is big and not fixed like DNA sequences. Experimental results (classification of bacteria rRNA and fungi rRNA) show the effectiveness of the proposed scheme. We will compare our method with other pattern classification method in future researches.

REFERENCES

[1] P. Baldi, S. Brunak, *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, Mass., 2001.

[2] R. A. Wallace, G. P. Sanders, and R. J. Ferl, *BIOLOGY: The Science of Life*, 3rd edition, HarperCollins Publishers Inc., 1991.

[3] I. Roitt, J. Brostoff, and D. Male, Immunology, 4th edition, Mosby, 1996.

[4] S. Forrest, A. S. Perelson, L. Allen, and R. Cherukuri "Self-nonself discrimination in a computer," *Proc. IEEE Symp. Research Security Privacy*, pp. 202–212, 1994.

[5] D. Dasgupta and S. Forrest, "An anomaly detection algorithm inspired by the immune system," in *Artificial Immune Systems and Their Applications*, D. Dasgupta, Ed. Springer, pp. 262–276, 1999.

[6] P. D'haeseleer, S. Forrest, and P. Helman, "An immunological approach to change detection: algorithms, analysis, and implications," *Proc. IEEE Symp. Computer Security Privacy*, pp. 110–119, 1996.

[7] J. Kim, and P. J. Bentley, "Towards an artificial immune system for network intrusion detection: An investigation of clonal selection with a negative selection operator," *Proc. Congr. Evolutionary Computation*, pp. 1244–1252, 2001.

[8] K. B. Sim and D. W. Lee, "Self-nonself recognition algorithm based on positive and negative selection," *IEICE Trans. on Info. & Syst.*, vol. E87–D, no. 2, Feb. 2004.

[9] F. Esponda, S. Forrest, and P. Helman, "Formal framework for positive and negative detection scheme," *IEEE Trans. Syst., Man, and Cybern., Part B*, vol. 34, no. 1, pp. 357–373, Feb. 2004.

[10] M. Gelfand, "Prediction of function in DNA sequence analysis," *J. Computational Biology*, vol. 1, pp. 87–115, 1995.

[11] K. B. Hwang, D. Y. Cho, S. W. Park, S. D. Kim, and B. T. Zhang, "Applying machine learning techniques to analysis of gene expression data: Cancer diagnosis," in *Methods of Microarray Data Analysis*, Kluwer Academic Publishers, pp. 167–182, 2002.

[12] G. B. Fogel, K. Chellapilla, and D. B. Fogel, "Identification of coding regions in DNA sequences using evolved neural networks," in *Evolutionary Computation in Bioinformatics*, G. B. Fogel and D. W. Corne, Eds., San Francisco, CA: Morgan Kaufmann, pp. 196–218, 2003.

[13] M. Ayara, and J. Timmis, *et al*. "Negative selection: How to generate detectors," *Proc. Int. Conf. ARtificial Immune System (ICARIS 2002)*, pp. 89–98, 2002.

[14] S. T. Wierzchon, "Generating optimal repertorie of antibody strings in an artificial immune system," in *Intelligent Information Systems*, M. Klopotek, M. Michalewicz and S. T. Wierzchon, Eds., *Advances in Soft computing Series of Physica-Verlag/Springer-Verlag*, pp. 119–133, 2000.

[15] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, pp. 189–192, Addison Wesley, 1989.

[16] Comparative RNA Web Site, http://www.rna.icmb. utexas.edu/