Modeling of Positive Selection for the Development of a Computer Immune System and a Self-Recognition Algorithm

Kwee-Bo Sim and Dong-Wook Lee

Abstract: The anomaly-detection algorithm based on negative selection of T cells is representative model among self-recognition methods and it has been applied to computer immune systems in recent years. In immune systems, T cells are produced through both positive and negative selection. Positive selection is the process used to determine a MHC receptor that recognizes self-molecules. Negative selection is the process used to determine an antigen receptor that recognizes antigen, or the nonself cell. In this paper, we propose a novel self-recognition algorithm based on the positive selection of T cells. We indicate the effectiveness of the proposed algorithm by change-detection simulation of some infected data obtained from cell changes and string changes in the self-file. We also compare the self-recognition algorithm based on positive selection with the anomaly-detection algorithm.

Keywords: Immune system, MHC set, negative selection, positive selection, self-recognition algorithm.

1. INTRODUCTON

The damage from computer viruses and hacking has been augmented with an increase in the use of computers and the Internet. In the same way that a virus invades to cause a disease, a computer virus is a program that invades a computer, damaging data files, destroying other profitable programs, and disturbing computer operations. It also copies itself to other computers through the Internet. Hacking is the skill of intruding into a computer for the purpose of extracting data and destroying the system. This type of destruction has also been on the rise. To protect computers against damage caused by hacking and viruses, research of the biological immune system for application in detecting intrusions [1-6] and viruses [7, 8] is in progress.

The immune system has a function that can discriminate between self (the normally occurring patterns in the system being protected e.g. body) and nonself (foreign pathogens, such as bacteria or viruses, or components of self that are no longer functioning normally). The representative immune cell is the cytotoxic T cell, which has a self-recognition component and an antigen receptor used to locate and eliminate infected cells [9, 10]. By modeling the characteristics of the biological immune system (BIS), the system that protects from damage by external attacks and eliminates intruders in the case of computer technology is called the computer immune system (or artificial immune system) [11, 12].

In this paper, we propose an algorithm that is modeled on the way in which living things discriminate between a self-cell and an antigen. An example of modeling on the self-recognition characteristics of the BIS is the anomaly-detection algorithm of Forrest *et al.* [7, 13, 14]. This algorithm utilizes the process of negative selection to produce an immune cell, form an anomaly detector, and apply itself to the self-recognition algorithm. It has the merit of being able to recognize various unknown antigens (modifications) through the preparation of sufficient anomaly detectors. As such, it has been applied to the computer immune system in recent years [1, 2, 11-15]. The anomaly-detection algorithm of Forrest *et al.* used the binary r-contiguous matching rule. In [16], to improve the efficiency of matching, Singh proposed the m-ary r-contiguous matching rule. In a general computer system, the basic unit is an 8 or 16 bit character variable. Therefore, m-ary r-contiguous matching rule is more effective. Therefore, we use 8-ary 2-contiguous matching rule in this paper.

However, the anomaly-detection algorithm has some drawbacks. It is poor at recognizing self-part

Manuscript received September 27, 2002; revised March 21, 2003; accepted August 7, 2003. Recommended by Editorial Board member You-Jip Won under the direction of Editor Chung Choo Chung.

Kwee-Bo Sim is with the School of Electrical and Electronic Engineering, Chung-Ang University, 221 Heukseok-dong, Dongjak-gu, Seoul 156-756, Korea (e-mail: kbsim @cau.ac.kr).

Dong-Wook Lee is with the Information and Telecommunication Research Institute, Chung-Ang University, 221 Heukseok-dong, Dongjak-gu, Seoul 156-756, Korea (e-mail: dwlee@wm.cau.ac.kr).

elimination and it has a lower recognition rate to block change (or block modification; meaning that a series of symbols is replaced by other symbols in a file) than point one (meaning that a symbol or a character is replaced by another symbol in the file) in the same change rate [7]. Hence, to improve this shortcoming, we propose a novel self-recognition algorithm using positive selection among the process for producing an immune cell in this paper. The effectiveness of the proposed algorithm is verified by the comparison of self-recognition rate for self-part modification at symbol level and at block level. We also compare the self-recognition algorithm based on positive selection with that based on the anomaly-detection algorithm.

2. BIOLOGICAL IMMUNE SYSTEM

The protection system of living creatures, the immune system, is a complex and sophisticated structure to protect cells and organs from various external organisms or proteins called antigens, such as pathogens, viruses and so on. The basic elements of the immune system are two types of lymphocytes, B cells (B lymphocytes) and T cells (T lymphocytes). B cells take part in humoral responses that secrete antibodies, and T cells take part in cell mediated immunity that stimulate or suppress cells concerned with immune response and kill infected self-cells [9, 10].

Immune cells use the Major Histocompatibility Complex (MHC) molecule to recognize self-cells. The protein that represents each characteristic also exists in the individual. It is called the MHC molecule. The part that recognizes the MHC molecule is located in the body of an immune cell. It is the MHC receptor. The immune cell uses the MHC receptor to judge between a self-cell and a nonself cell. The immune cell such as a B cell or a T cell has a detector that recognizes specific antigen. This is known as the "antigenic receptor." A T cell has both the MHC receptor and an antigenic receptor [9].

2.1. The principle for developing an immune cell

In the BIS, an immune cell, which is the core of the immune response, relies on two elements to eliminate the antigens that have invaded a living body. One is cooperation and communication between cells. The other is the ability to discriminate between a self-cell and a nonself-cell. A representative immune cell is the cytotoxic T cell that has both an antigenic receptor to recognize the antigens and the MHC receptor to recognize the MHC molecule (MHC protein) that identifies a self-cell. A cytotoxic T cell is produced through a positive section and a negative one. If a T cell receptor does not operate correctly in the immune system, it recognizes a self-cell as an antigen and attacks it. Therefore, when T cells are produced initially, they are examined for the correct operation of the MHC receptor and antigenic receptor. This process is positive selection and negative one. Theses processes determine whether two receptors are operated correctly or not.

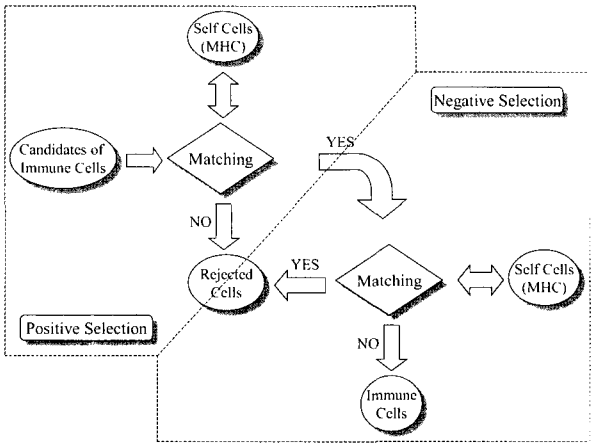

Fig. 1. Developing process of T cell.

Positive selection is a method used to examine the MHC recognition function of each immature immune cell since only immature immune cells able to recognize MHC molecules correctly in the self-cell can be used in an immune system. Mature immune cells consist of only cells in which the MHC receptor is matched with MHC molecules among the immature immune cells. At this time, the immune system can be maintained by elimination of the unmatched cells, because the unmatched immature immune cells cannot recognize a self-cell.

Negative selection is a method employed to exclude the immature immune cells that recognize a self-cell as an antigen. If an antigenic receptor recognizes a MHC molecule as an antigen, the antigenic receptor regards all self-cells as antigens. When immature immune cells conjoin to the MHC molecule, only those cells that the antigenic receptor does not recognize MHC molecules as antigens are selected. If an immature immune cell recognizes a MHC molecule as an antigen, it is eliminated.

The immature immune cells form an appropriate immune response in a living object after completing these two selections. The developing process for a mature immune cell is shown in Fig. 1.

3. SELF-NONSELF RECOGNITION ALGORITHMS

One of the most important characteristics of BIS is discrimination ability between self and nonself by recognizing a self-cell from an antigen. Forrest *et al.* proposed the anomaly-detection algorithm based on negative selection, which is one of the producing principles for immune cells [7, 13, 14]. This is an

algorithm that recognizes a nonself using antigen receptor and it has superior characteristics to detect local modification and addition of self-space. However, it has a lower self-recognition rate regarding self-part elimination [7] and block modification than point one. As such, in this paper, to improve the recognition rate in various conditions we propose an MHC detection algorithm based on positive selection of the BIS.

3.1. Anomaly-detection algorithm based on negative selection

The anomaly-detection algorithm based on negative selection is one of self-nonself discrimination algorithm proposed by Forrest *et al.* [7]. They composed a set of detectors that do not recognize self-space. A composed detector set is used for nonself recognition. This algorithm is divided into two sections. One is to compose the anomaly detectors by negative selection and the other is to check the occurrence of modifications by using the composed detector set. Fig. 2 represents the process to produce the anomaly detector set by negative selection. The anomaly detector consists of strings that do not correspond to self-space. To begin, define a self-space S to be protected. Then match the strings of S after making a set of random strings, R_0 followed by r-contiguous matching between each string in R_0 and all the strings in S. We can compose detector set, R, which does not match any string in S, where $r < n$. At this time, matched string set E is rejected.

A perfect match between two strings of the same length indicates that the symbols of each cell located in each position on the string are identical. Because it is difficult to locate an unmatched string as they become larger, a partial matching rule is used. The matching rule used by the anomaly-detection algorithm is an r-contiguous matching rule. If the same r-contiguous cells are located in the two strings, it is defined as being matched.

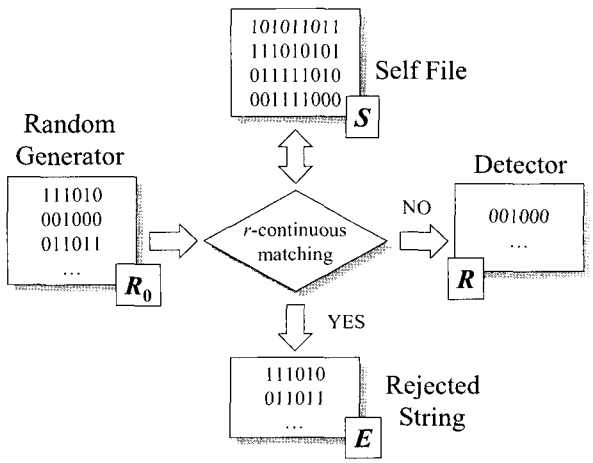

Fig. 2. Construction method for an anomaly detector.

We can recognize self and nonself using the anomaly detectors that were made by the above process. This algorithm has the merit that it is able to recognize various antigens, modifications, by preparing sufficient anomaly detectors. However, because this algorithm detects self-change by recognizing nonself, it is inefficient at detecting the partial elimination of self-space. Furthermore, it shows a lower recognition rate to block modification than point one in the same change rate. This result will be shown in the simulation results.

3.2. Self-recognition algorithm based on positive selection

In this section, we propose a novel algorithm based on positive selection to improve the recognition rate to block modification. This algorithm is a method to recognize self by modeling of the MHC detector, which is another receptor of T-cell. The proposed algorithm produces MHC detectors that have specific characteristics or some component of self-space to assist in recognizing self. The set of MHC detectors is called "MHC set" in this paper. Because MHC set has the characteristic of self-space, it can discriminate between self and nonself. The composing process of detector set using characteristics of self-space is positive selection, which is the process used to produce and examine the MHC receptor when T cells are produced initially.

If checking space has all components of the MHC set, it is recognized as self-space. Otherwise, it is considered nonself space or modified self-space. Fig. 3 shows the elements of MHC detector string (MHC string). Each string location is a cell, where each cell is an 8 bit character variable. As such, it is one of the 256 symbols. The code, basic unit of matching, refers to the r-contiguous cell. The MHC detector is produced from the self-space, S. The MHC string is composed of partially matched codes.

Detailed algorithm of constructing a MHC set is as follows.

Step 1: Define a self-space S and divide S into strings of fixed length.

Self-space is divided into the strings of fixed length ($= l$). Thus, self-space becomes the set of strings.

Step 2: Select the first code of MHC string.

Select the first code that is not equal to that of other MHC strings among the codes to occupy the first location of all strings in the self-space. This code

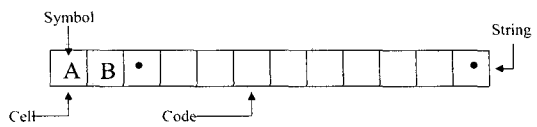

Fig. 3. Elements of MHC detector string.

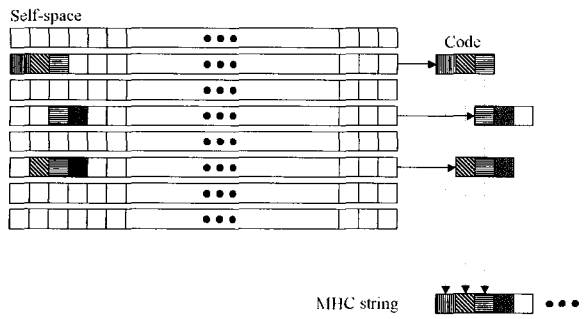

Fig. 4. Construction method for a MHC string.

is the seed to create the MHC string.

Step 3: Select the next code.

Select another code that identical to the second and later cells of the first code. The selected code is the code for the MHC string in the second location.

Step 4: Determine the last code.

Continue to select codes in the same way. If partially matched codes are not in the self-space strings, go to Step 2 and determine a new first code.

Step 5: Compose a MHC set that is a predefined number of the MHC string.

Until the predefined number of the MHC strings is composed, repeat from Step 2 to Step 4.

Fig. 4 shows the construction method for an MHC string.

The checks for self-recognition are operated using the MHC set composed. The codes of each MHC string are examined to confirm whether they occupy a location in the self-space string. If all codes of the MHC string are present in each location of the checked space, the verified space is recognized as self by that MHC string. Otherwise, it is recognized as nonself. Moreover, if all codes in an MHC set recognize the checked space as self, then the checked space is recognized as self-space. Fig. 5 shows the process of self-nonself discrimination using the MHC set composed.

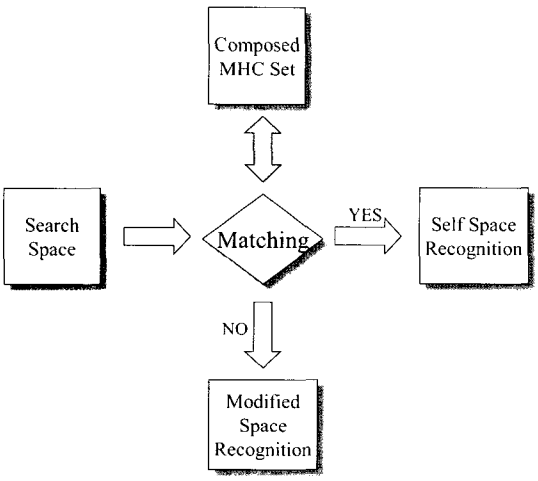

Fig. 5. Process of self-nonself discrimination using the MHC set composed.

The proposed MHC-set algorithm has some characteristics as follows. Firstly, as we determine the cells of the MHC string, the MHC string memorizes position information. Therefore, it can easily detect the elimination or addition of various strings in the search-space. In addition, distributed codes of the MHC string increase recognition rate to block modification. Secondly, the MHC string is composed of a code that has duplicated cells. As such, each cell of the MHC string represents the duplicated position as many as the length of the code ($= r$). For example, in Fig. 4, the length of the code is 3 ($r = 3$); therefore each cell represents the cells of 3 positions.

4. SIMULATION RESULTS

To compare the MHC-set algorithm with the anomaly-detection algorithm we obtained the change detection rate of point and block modification of the self-space by simulation. The self-space is a set of strings, where a string was composed of 32 cells ($l = 32$), and a cell was an 8 bit character variable. In simulation, the number of self-space strings was 800, 1600, and 3200. The number of MHC strings (and anomaly detector) was set at 10 and 20. The code was set at 2 for matching and producing ($r = 2$). The length of the anomaly detector was identical to that of the MHC string. It is set at 32. Also, we used the 8-ary 2-contiguous matching rule. Two methods were used for the method of modification of self-space. One is used in the cell change method to modify certain cells in the self-space in order to verify the self-recognition rate for the point modification of self-space occurring due to noise. The other is used in the string change method to modify all cells of a string in the self-space in order to certify self-

Table 1. Recognition rate for cell modification (No. of detectors = 10).

No. of detectors		Size of self-space (No. of strings)					
10		800		1600		3200	
		MHC set	ADs	MHC set	ADs	MHC set	ADs
Change rate	0.01	18.04	8.06	17.45	14.87	18.58	26.8
	0.02	31.46	14.98	31.78	26.53	33.15	44.94
	0.03	43.18	20.75	45.06	35.44	44.6	60.05
	0.04	53.38	26.53	54.95	44.27	54.82	69.44
	0.05	62.44	31.95	61.51	52.14	61.71	77.42
	0.06	69.15	35.34	69.74	58.46	70.28	82.57
	0.07	73.86	40.07	75.64	64.16	74.19	86.91
	0.08	76.74	43.2	78.64	68.02	80.62	90.01
	0.09	81.75	46.34	82.06	70.47	81.54	92.52
	0.10	86.29	49.31	86.42	74.58	85.87	94.05

Table 2. Recognition rate for cell modification (No. of detectors = 20).

No. of detectors		Size of self-space (No. of strings)					
20		800		1600		3200	
		MHC set	ADs	MHC set	ADs	MHC set	ADs
Change rate	0.01	31.73	13.99	31.87	26.34	32.57	46.15
	0.02	49.75	27.47	54.16	44.69	54.24	70.17
	0.03	68.54	36.1	68.59	59.79	69.98	83.69
	0.04	85.38	52.32	77.63	69.71	78.34	90.28
	0.05	84.18	52.04	85.55	77.09	85.04	94.92
	0.06	90.46	58.18	90.4	82.63	89.53	96.93
	0.07	92.68	63.78	92.7	86.6	94.25	98.19
	0.08	95.18	67.94	95.29	89.73	96.15	98.83
	0.09	96.56	72.46	97.27	92.2	97.44	99.52
	0.10	97.89	74.95	97.57	93.77	97.52	99.65

Table 3. Recognition rate for string modification (No. of detectors = 10).

No. of detectors		Size of self-space (No. of strings)					
10		800		1600		3200	
		MHC set	ADs	MHC set	ADs	MHC set	ADs
Change rate	0.01	21.34	3.59	34.9	7.1	63.12	14.46
	0.02	58.55	7.23	48.25	14.56	85.04	27.3
	0.03	45.57	10.92	84.98	21.56	96.7	37.56
	0.04	64.7	14.33	85.01	26.51	93.3	46.31
	0.05	69.52	17.5	91.99	32.2	96.82	53.45
	0.06	79.51	21.77	95.64	37.04	97.44	60.57
	0.07	85.36	22.81	98.99	41.88	99.97	66.16
	0.08	88.46	26.39	99.98	46.4	100	72.19
	0.09	92.88	29.23	99.98	50.16	100	75.5
	0.10	97.75	32.17	99.98	54.51	100	79.43

recognition rate for block modification of self-space occurring due to hacking or to a computer virus. Each simulation was repeated 10,000 times for stochastic reliance.

We obtained the recognition rate for cell modification of the self-space. At this time, the change rate varied from 0.01 to 0.1. Tables 1 and 2 show the recognition rate for cell modification, when the number of detectors is 10 and 20 each.

As a result, the MHC-set algorithm represents robust performance against the size of the self-space. However, in the anomaly-detection algorithm, as the size of the self-space increases, the recognition rate also increases. Therefore, when the size of the self-space is small, the MHC-set algorithm is more effective than the anomaly-detection algorithm. On the contrary, when the size of the self-space is large,

Table 4. Recognition rate for string modification (No. of detectors = 20).

No. of detectors		Size of self-space (No. of strings)					
20		800		1600		3200	
		MHC set	ADs	MHC set	ADs	MHC set	ADs
Change rate	0.01	44.07	7.17	62.84	14.58	86.51	27.27
	0.02	57.98	14.78	77.31	27.23	97.73	46.72
	0.03	69.53	20.85	94.87	37.17	99.23	60.92
	0.04	90.33	31.92	100	46.08	100	71.97
	0.05	93.12	31.41	100	54.53	100	78.64
	0.06	99.23	36.19	100	60.77	100	84.8
	0.07	99.98	42.02	100	66.52	100	88.46
	0.08	99.99	46.45	100	71.13	100	91.19
	0.09	100	50.54	100	75.31	100	94.16
	0.10	100	53.77	100	78.99	100	95.58

the anomaly-detection algorithm shows good performance.

Tables 3 and 4 show the results of recognition rate for string modification of the self-space. The recognition rate for string modification in the MHC-set algorithm is higher than that in the anomaly-detection algorithm. This means that the MHC-set algorithm, which has the space for self-characteristics, is better at recognizing modifications of self-space produced by a change in the block.

As the results indicate, the MHC-set algorithm has robust performance against the size of the self-space in cell and string modification. In particular, the MHC-set algorithm shows superior results in string modification.

5. CONCLUSIONS

In this paper we proposed the self-recognition algorithm based on the computer's ability to recognize self-space, modeled on the MHC molecule. This is the MHC-set algorithm. We also compared it with the anomaly-detection algorithm, which is based on negative selection. In simulation results, the MHC-set algorithm shows a higher self-recognition rate when detecting block (string) changes. It also has a higher self-recognition rate when detecting point (cell) changes, although, it has a lower self-recognition rate for a few conditions. In the future, if the two complementary algorithms are combined, it is expected that they can be used as a basic algorithm for an innovative computer immune system.

REFERENCES

[1] S. Hofmeyr, S. Forrest, and A. Somayaji, "Intrusion detection using sequences of system calls," *Journal of Computer Security*, vol. 6, pp.

151-180, 1998.

[2] C. Warrender, S. Forrest, and B. Pearlmutter, "Detecting intrusions using system calls: Alternative data models," *Proc. of the IEEE Symposium on security and Privacy*, pp. 133-145, 1999.

[3] D. Dasgupta, "An immune agent architecture for intrusion detection," *Proc. of the Genetic and Evolutionary Computation Conference (GECCO 2000) Workshop Program*, pp. 42-44, 2000.

[4] J. Gu, D. Lee, S. Park, and K. Sim, "An immunity-based security layer model," *Proc. of the Genetic and Evolutionary Computation Conference (GECCO 2000) Workshop Program*, pp. 47-48, 2000.

[5] J. Gu, D. Lee, K. Sim, and S. Park, "An antibody layer for internet security," *Proc. of the Global Telecommunication Conference (GLOBECOM 2000)*, pp. 450-454, 2000.

[6] J. Gu, D. Lee, K. Sim, and S. Park, "An immunity-based security layer against Internet antigens," *Trans. on IEICE*, vol. E83-B, no.11, pp. 2570-2575, 2000.

[7] S. Forrest, A.S. Perelson, L. Allen, and R. Cherukuri, "Self-nonself discrimination in a computer," *Proc. of the IEEE Symposium on Research in Security and Privacy*, pp. 202-212, 1994.

[8] P. Harmer, and G. Lamont, "An agent based architecture for a computer virus immune system," *Proc. of the Genetic and Evolutionary Computation Conference (GECCO 2000) Workshop Program*, pp. 45-46, 2000.

[9] I. Roitt, J. Brostoff, and D. Male, *Immunology*, 4th edn., Mosby, 1996.

[10] R. A. Wallace, G. P. Sanders, and R. J. Ferl, *Biology: The science of life*, 3rd edn., HarperCollins, 1991.

[11] A. Somayaji, S. Hofmeyr, and S. Forrest, "Principles of a computer immune system," *Proc. of the New Security Paradigms Workshop*, pp. 75-82, 1998.

[12] S. Forrest, S. Hofmeyr, and A. Somayaji. "Computer immunology," *Communications of the ACM*, vol. 40, no. 10, pp. 88-96, 1997.

[13] P. D'haeseleer, S. Forrest, and P. Helman, "An immunological approach to change detection: algorithms, analysis, and implications," *Proc. of the IEEE Symposium on Computer Security and Privacy*, pp. 110-119, 1996.

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

[15] M. Ayara, J. Timmis, R. de Lemos, L. de Castro, and R. Duncan, "Negative selection: how to generate detectors," *Proc. of 1st International Conference on Artificial Immune System*, pp. 89-98, 2002.

[16] S. Singh, "Anomaly detection using negative selection based on the r-contiguous matching rule," *Proc. of 1st International Conference on Artificial Immune System*, pp. 99-106, 2002.

Kwee-Bo Sim received the B.S. and M.S. degrees in the Department of Electronic Engineering from Chung-Ang University in 1984 and 1986 respectively, and Ph.D. degree in the Department of Electrical Engineering from The University of Tokyo, Japan, in 1990. Since 1991, he has been a faculty member of the School of Electrical and Electronic Engineering at Chung-Ang University, where he is currently a Professor. His areas of interest include artificial life, neuro-fuzzy and soft computing, evolutionary computation, learning and adaptation algorithms, autonomous decentralized systems, intelligent control and robot systems, artificial immune systems, evolvable hardware, and artificial brain etc. He is a member of IEEE, SICE, RSJ, KITE, KIEE, ICASE, and KFIS.

Dong-Wook Lee received the B.S., M.S., and Ph.D. degrees in the Department of Control and Instrumentation Engineering from Chung-Ang University in 1996, 1998, and 2000, respectively. Since 2002, he has been with the Information and Telecommunication Research Institute at Chung-Ang University, where he is currently a Research Professor. His areas of interest include artificial life, evolutionary computation, artificial brain, and artificial immune systems.