생체분자 퍼셉트론의 신뢰성 향상을 위한 열역학 기반 가중치 코딩 방법

(Thermodynamics-Based Weight Encoding Methods for Improving Reliability of Biomolecular Perceptrons)

임 희 웅 † 유 석 인 †† 장 병 탁 ††

(Hee-Woong Lim) (Suk I. Yoo) (Byoung-Tak Zhang)

요 약 생체분자 컴퓨팅은 DNA와 같은 생체 분자를 이용하여 정보를 표현하고 처리하는 새로운 컴퓨팅 패러다임이다. 작은 부피에 존재하는 무수히 많은 분자와 화학 반응에 내재된 대규모 병렬성은 새로운 개념의 고성능 계산 기법에 영감을 주었고 이를 바탕으로 다양한 계산 모델 및 문제 해결을 위한 분자 알고리즘이 개발되었다. 한편 생체 분자를 이용한 정보처리라는 특징은 생물학 문제에 적용될 수 있는 가능성을 시사한다. 유전자 발현 패턴과 같은 생화학적 분자 정보의 분석을 위한 도구로서의 가능성을 가지고 있는 것이다. 이러한 맥락에서 DNA 컴퓨팅 기반의 생체분자 퍼셉트론 모델이 제안되었고 그 실험적 구현 결과가 제시된 바 있다. 생체분자 퍼셉트론의 핵심인 가중치 표현 및 가중치-합 연산은 입력 분자와 가중치를 표현하는 프로브 분자간의 경쟁적 혼성화 반응에 기반하고 있다. 그러나 그 혼성화 반응에서 열역학적 대칭성을 가정하고 있기 때문에 사용하는 프로브에 따라 가중치 표현의 오차가 있을 수 있다. 본 논문에서는 비대칭적인 열역학적 특성을 고려하여 일반화된 혼성화 반응 모델을 제시하고, 이를 바탕으로 신뢰성 있는 생체 분자 퍼셉트론의 구현을 위한 가중치 코딩 방법을 제안한다. 그리고 본 논문에서 제시한 가중치 표현 방법의 정확성을 이전 모델과 컴퓨터 시뮬레이션을 통해 비교하고 한계 오차를 만족하기 위한 조건을 제시한다.

키워드 : DNA 컴퓨팅, 생체분자 퍼셉트론, 가중치 표현, 혼성화 반응 모델

Abstract Biomolecular computing is a new computing paradigm that uses biomolecules such as DNA for information representation and processing. The huge number of molecules in a small volume and the innate massive parallelism inspired a novel computation method, and various computation models and molecular algorithms were developed for problem solving. In the meantime, the use of biomolecules for information processing supports the possibility of DNA computing as an application for biological problems. It has the potential as an analysis tool for biochemical information such as gene expression patterns. In this context, a DNA computing-based model of a biomolecular perceptron has been proposed and the result of its experimental implementation was presented previously. The weight encoding and weighted sum operation, which are the main components of a biomolecular perceptron, are based on the competitive hybridization reactions between the input molecules and weight-encoding probe molecules. However, thermodynamic symmetry in the competitive hybridizations is assumed, so there can be some error in the weight representation depending on the probe species in use. Here we suggest a generalized model of hybridization reactions considering the asymmetric thermodynamics in competitive hybridizations and present a weight encoding method for

· 본 연구는 교육인적자원부 BK21-IT 및 산업자원부 차세대 신기술 개발 사업의 분자 진화 컴퓨팅(MEC) 과제에 의하여 일부 지원되었다. 또한 이 연구를 위해 장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에도 감사 드린다.

† 정 회 원 : 서울대학교 전기컴퓨터공학부
hwlim@ailab.snu.ac.kr
†† 종신회원 : 서울대학교 전기컴퓨터공학부 교수
siyoo@ailab.snu.ac.kr
btzhang@bi.snu.ac.kr

논문접수 : 2006년 10월 26일
심사완료 : 2007년 10월 18일

정보과학회논문지: 소프트웨어 및 응용 제34권 제12호(2007.12)

the reliable implementation of a biomolecular perceptron based on this model. We compare the accuracy of our weight encoding method with that of the previous one via computer simulations and present the condition of probe composition to satisfy the error limit.

Key words : DNA computing, biomolecular perceptron, weight encoding, hybridization reaction model

1. 서 론

최근 생체분자와 이에 대한 생화학적인 실험 과정을 통해 정보를 표현, 저장, 그리고 처리하는 생체분자 컴퓨팅(biomolecular computing)이 새로운 계산 패러다임으로 대두되었다. 특히 DNA는 그것을 구성하는 기본 단위들 간의 특이적인 혼성화 반응이라는 특성으로 인해 새로운 정보 표현의 매개체로서의 가능성을 제시하였으며 NP 문제 해결을 위한 분자 알고리즘과 대용량 연관 메모리를 포함한 다양한 응용 방법에 대한 연구가 이루어졌다[1-6]. 그러나 정보처리에 사용되는 DNA 염기서열 디자인이나 계산 과정으로서의 생화학 반응의 제어와 같은 현실적인 문제에 부딪혀 고성능/고용량 정보처리 방법으로서의 DNA 컴퓨팅은 아직 작은 크기의 문제에 국한되어 있다. 그래서 이러한 한계를 극복하기 위해 효과적인 정보 표현을 위한 DNA 염기 서열 디자인 방법들이 제안되었으며[7-9], 이와 더불어 DNA를 이용한 논리 게이트, 분자 오토마타, 행렬 및 벡터 연산을 위한 분자 알고리즘을 비롯한 다양한 계산 모델들이 개발되었다[10-12].

이와 같이 DNA나 효소 등과 같은 생화학 물질들이 정보 처리 및 계산을 위한 매개체로서의 역할을 할 수 있다는 사실은 DNA 컴퓨팅의 생물학 문제에 대한 응용으로서의 가능성을 내포하고 있다. 최근에 발표된 지능형 DNA 칩[13], 유전자 발현 분석을 위한 방법[14], 그리고 유전자 발현의 논리적 조절을 위한 분자 오토마타[15]는 이러한 생물학 문제에 적용된 DNA 컴퓨팅의 예라 할 수 있다. 정보처리의 관점에서 볼 때, 유전자 발현 패턴의 분석이나 이에 기반한 질병 진단과 같은 작업들은 생화학 분자의 형태로 존재하는 패턴이 마이크로어레이[16] 방법과 같은 다양한 측정 방법을 통해 전기적 신호의 형태로 변환된 뒤 컴퓨터상에서(in silico) 다양한 생물정보학 패턴 분석 방법들을 이용해서 처리되는 과정으로 생각할 수 있다. 반면 DNA 컴퓨팅은 그러한 정보의 변환 과정이 없이 생화학 물질로 표현 분자 패턴을 시험관 내에서(in vitro) 직접 처리하기 때문에 변환 과정에서의 정보 손실을 줄이고 대상 패턴의 크기에 관계 없이 병렬적인 연산을 통해 시간적으로나 경제적으로 효율적인 문제 해결을 가능하게 하는 것이다.

이러한 생화학적 분자 패턴 분석 기술의 구현을 위해서는 다양한 대수적 연산들을 생화학적 실험 과정을 통해서 구현하는 분자 알고리즘의 개발이 필요하다. 이러한 맥락에서 DNA나 RNA의 형태로 수치정보를 표현하고 이에 대한 벡터나 행렬 연산을 위한 DNA 컴퓨팅 기반의 방법들이 Oliver[17]나 Mills[18]에 의해 제안된 바 있으나 구현을 위한 실험 절차가 복잡하여 확장성이 부족하다는 단점이 있었다. 이러한 한계를 극복하기 위해 범용의 행렬 연산대신 목표 분자 패턴의 분석에 필요한 기본 연산으로서 가중치-합 연산을 정의하고, 단순한 절차로 구현되는 DNA 기반의 가중치-합 연산을 위한 분자 알고리즘을 개발하여 이를 바탕으로 분자 패턴의 분류를 위한 생체분자 퍼셉트론이 실험적으로 구현된 바 있다[19].

퍼셉트론은 인공신경망을 구성하는 기본 단위로서 실수 벡터로 표현된 입력 벡터의 이진 분류를 수행할 수 있다. 이러한 퍼셉트론을 DNA를 이용한 생화학 실험 과정으로 구현하여 DNA혹은 RNA용액으로 표현된 입력 패턴의 분류를 수행가능 하도록 만든 것이 바로 생체분자 퍼셉트론이다. 입력 패턴의 각 요소들은 DNA 기반의 가중치-합 연산을 통해 해당 패턴의 분류 결과를 나타내는 지시 정보를 만들어내는데 사용된다. 구체적으로, 가중치-합 연산은 입력 패턴을 표현한 입력 DNA 가닥과, 가중치를 표현하는 서로 다르게 라벨링 되어있는 한 쌍의 프로브 가닥들 사이의 상보성에 기반한 경쟁적인 혼성화 반응에 의해 이루어진다. 이 때, 가중치의 표현은 이 경쟁적인 두 혼성화 반응간의 열역학적 대칭이라는 가정하에 연산 과정에서 사용되는 프로브의 혼합 비율에 의해 이루어진다. 그러나 실제 구현의 측면에서는 사용하는 프로브 DNA의 구성 혹은 라벨링된 물질에 따라 혼성화 반응의 열역학적 특성이 달라져 가중치 표현에 오차가 있을 수 있다. 이에 본 논문에서는 가중치 표현을 위한 경쟁적 혼성화 반응 모델을 열역학적인 측면에서 비대칭적인 반응으로 확장하여 이를 바탕으로 신뢰성 있는 가중치 표현 방법을 제시하고 다양한 조건하에서의 시뮬레이션을 통해 기존의 가중치 표현 방법과 비교 분석한다.

다음 장에서는 DNA 컴퓨팅 기반 생체분자 퍼셉트론의 개념 및 가중치-합 연산을 위한 분자 알고리즘에 대해서 간략히 설명하며, 3장에서는 비대칭적인 열역학적

특성을 고려한 일반화된 혼성화 반응모델을 세우고 이를 바탕으로 향상된 가중치 표현 방법을 제시한다. 4장에서는 이 모델을 바탕으로 한 컴퓨터 시뮬레이션 결과와 효과적인 가중치 코딩 방법 및 조건을 제시하고, 마지막으로 5장에서 결론을 맺는다.

2. DNA 기반 생체분자 퍼셉트론

패턴 분류 문제는 어떤 데이타나 패턴을 입력으로 받아 이미 알고 있는 기본 지식과 해당 입력으로부터 추출된 통계적 특징을 바탕으로 입력 패턴을 분류하는 작업을 말한다. 패턴 분류 문제는 로봇비전, 스팸필터링, 문자인식과 같이 여러 분야에서 찾아 볼 수 있으며 다양한 방법론이 연구되고 개발되어 왔다. 이러한 문제들은 생물학 분야에서도 찾아볼 수 있다. 게놈프로젝트의 완료 후 생명 현상에 있어서의 다양한 유전자들의 역할에 대한 연구가 활발하게 이루어져 왔으며, 특히 질병에 관련된 유전자와 표지 물질의 발견 및 분석은 중요한 이슈로 대두되었다. 이러한 생물학 문제들은 생체분자로 표현된 패턴 분류 문제로 해석할 수 있는데, 특히 표지 물질에 해당하는 messenger RNA(mRNA)나 단백질의 정량적 정보들은 그 패턴 분류를 위한 유용한 특징에 해당한다. 기존의 생물정보학이 이러한 특징을 컴퓨터 상의 수치데이타로 표현하여 처리한 반면, DNA 기반의 생체분자 퍼셉트론 해당 시험관 내의 분자 패턴을 직접 처리하여 분류를 수행한다[19]. 여기에서는 [19]에서 제시되었던 DNA 기반의 생체분자 퍼셉트론의 개념적 모델을 기존의 퍼셉트론과 비교하여 설명하고 이를 생화학적 실험 과정으로 구현하기 위한 분자 알고리즘을 요약 설명한다.

2.1 개념적 모델

일반적인 퍼셉트론[20,21]은 인공신경망을 구성하는 기본 단위로서, 실수 벡터를 입력 패턴으로 받아 각 입력 요소에 할당된 가중치를 고려하여 가중치-합을 계산하고 결과를 기준 값과 비교하여 입력 벡터의 분류 결과를 나타내는 최종 출력 값을 내놓는다(그림 1).

DNA 기반의 생체분자 퍼셉트론 또한 마찬가지로 실수 벡터를 입력으로 받아 분류를 수행한다. 그러나 목표 패턴이 생체분자로 표현된 정보라는 것과 분류를 위한 핵심 연산 과정이 시험관 수준에서 생화학적 실험 과정을 통해 수행된다는 점에서 다르며, 따라서 그 구현을 위해 아키텍처 면에서 몇 가지 차이가 있다. 먼저 기존의 퍼셉트론이 부호와 관계 없이 실수 벡터를 입력으로 취하는 반면, 생체분자 퍼셉트론은 0 혹은 그 이상의 값으로만 구성된 실수 벡터를 입력으로 받아들인다. 이는 뒤에 자세히 언급되겠지만 입력 패턴이 분자의 정량적 정보에 의해 표현되고 이러한 정보는 언제나 0 이상의 값을 가지기 때문인데, 특정 유전자의 발현 수준을 나타내는 mRNA의 양이 그 대표적인 예이다. DNA 기반의 퍼셉트론은 유전자 발현 패턴과 같은 생화학 분자의 정량적 정보의 분류를 주 목표로 하고 있기 때문에 입력 패턴에 대한 이러한 제약은 크게 문제 되지 않는다. 둘째로, 기존 퍼셉트론은 입력에 대한 가중치-합 연산을 통해 분류 결과를 나타내는 단일 출력을 만들어 내는 반면, DNA 기반의 퍼셉트론은 두 개의 서로 상보적인 출력을 만들어 내며, 이 두 값은 입력 패턴의 분류를 위한 지시 정보로서 서로 비교하여 최종 분류 결과를 결

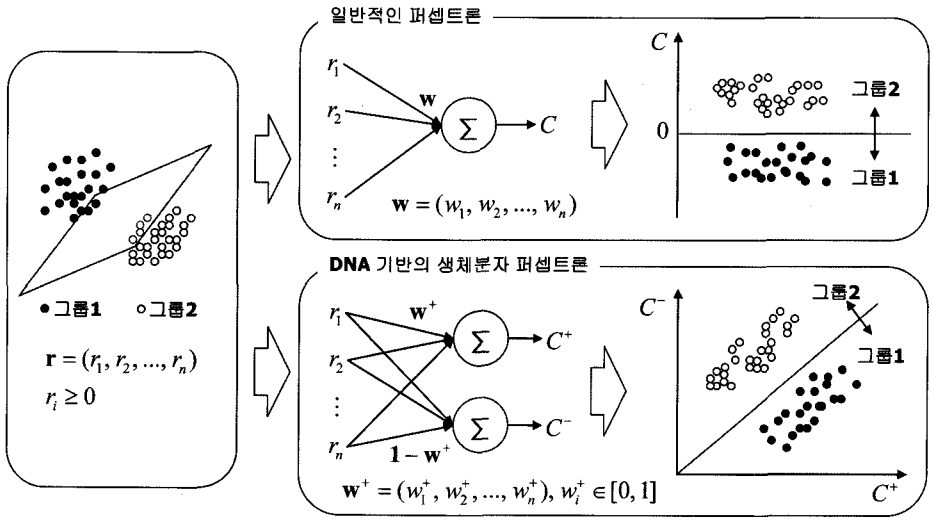


그림 1 일반적인 퍼셉트론과 DNA 기반의 생체분자 퍼셉트론의 개념도

정하는데 사용된다(그림 1). 이와 같이 두 개의 상보적인 출력을 만들어 내는 것은 상기 언급한 입력과는 달리 패턴의 성공적인 분류를 위해서는 음의 값을 가지는 가중치 또한 필요하기 때문이다. 그러나 앞서 말한 바와 같이 DNA를 비롯한 분자의 양적 정보는 항상 0 이상의 값을 가지기 때문에, 음의 가중치 값을 명시적으로 표현하여 단일 출력 값을 계산하는 대신 두 개의 서로 상보적인 출력 값을 만들어 내는 것이다.

구체적으로 살펴보면, 그림 1에 보이는 바와 같이 DNA 기반의 퍼셉트론 $f(\mathbf{r})$은 실수 벡터 $\mathbf{r} = (r_1, r_2, ..., r_n)$을 입력으로 받아 각 입력 요소에 할당된 가중치 w^+를 이용하여 두 개의 가중치-합 결과

$$C^+ = \sum_{i=1}^{n} r_i w_i^+,\ C^- = \sum_{i=1}^{n} r_i (1 - w_i^+)$$

를 계산하여 출력한다. 이 두 출력은 서로 비교하여 다음과 같이 최종 분류 결과를 출력하는데 사용된다.

$$f(\mathbf{r}) = \begin{cases} \text{if } C^+ \geq C^-, \text{ then } \mathbf{r} \in \text{그룹1} \\ \text{if } C^+ < C^-, \text{ then } \mathbf{r} \in \text{그룹2} \end{cases}$$

이러한 일련의 작업은 n차원 공간의 한 점으로 표현되는 입력 패턴을 가중치-합 연산을 통해 2차원 평면의 한 점으로 변환하여 패턴의 분류를 수행하는 과정으로 생각할 수 있다. DNA 기반 퍼셉트론은 계산 복잡도의 측면에서 그리 복잡하진 않고 컴퓨터상에서 구현되는 모델 가운데 극히 단순한 모델에 속한다. 그러나 여기에서는 생화학적 분자 패턴에 대한 시험관 내 직접 연산에 초점을 맞추어, 패턴 분류를 위한 기본 연산으로서 가중치-합 연산을 정의하고 확장성 및 실용성을 고려한 구현을 목표로 하고 있다. 이러한 패턴 분류 과정은 유전자 기반 진단의 관점에서 볼 때, 유전자 발현 패턴을 구성하는 표지 유전자의 mRNA와 그 정량 정보로 구성된 n차원 입력 벡터에 대해, 각 유전자 마다 개별적인 가중치를 가지고 발현 수준을 종합하여 질병의 유무 혹은 질병의 분류를 수행하는 과정으로 생각 할 수 있다.

2.2 분자 알고리즘

DNA 컴퓨팅에서의 분자 알고리즘은 일종의 아날로그 컴퓨팅으로서 DNA 분자들 간의 결합이나 효소 반응과 같은 조작 과정을 기호적인 정보 혹은 수치 정보의 처리 과정으로 대응시킨 것을 말한다. DNA 분자들 간의 상보적인 혼성화 반응과 화학 반응에 내재된 병렬성은 염기 서열을 이용한 기호 정보의 표현과 연산에서의 병렬 처리 방법을 제공하는데, 초기의 NP 문제 해결을 위한 분자 알고리즘들이 그 예라 할 수 있다[4-6]. 뿐만 아니라 DNA 및 생화학 분자의 정량적 정보와 수치 정보 사이의 유사성은 수치 정보의 표현 및 연산을 위한 분자 알고리즘의 바탕을 제공하였다[17,18].

DNA 기반의 퍼셉트론은 입력 패턴의 처리를 위해 가중치-합 연산을 기본 연산자를 사용하는데, 이는 퍼셉트론 뿐만 아니라 다른 여러 패턴 인식 방법에서도 사용되는 기본 연산자이다. 이러한 가중치-합 연산의 시험관 내 구현을 위해서는 DNA 분자를 이용한 입력 벡터의 표현과 각 입력에 대한 가중치 부여 방법이 필요하다. 기본적으로 n차원 실수 입력 벡터 $\mathbf{r} = (r_1, r_2, \cdots, r_n)$은 서로 다른 고유의 염기서열을 가진 n개의 DNA 가닥 R_i (i = 1, 2, ..., n)에 의해 표현되고, 입력 벡터의 각 요소 r_i는 해당 입력 가닥 R_i의 양에 의해 표현된다. 따라서 앞서 언급한 바와 같이 0 이상의 실수로 이루어진 입력 벡터를 표현 가능하다. 한편 입력 요소에 대응되는 가중치 w_i^+는 각 입력 가닥에 특이적으로 결합하는 한 쌍의 프로브, 리포터 P_i와 경쟁자 P_i^*의 혼합 비율로 표현된다. 여기에서 프로브란 특정 목표 DNA 분자의 양을 측정할 수 있도록 도와주는 DNA로서, 해당 목표 DNA에 특이적으로 결합할 수 있도록 목표 DNA에 상보적인 서열로 디자인되어있으며, 특정 파장대의 형광신호를 발산하는 물질을 한쪽 끝에 부착하여 해당 파장대에서 방출되는 신호의 강도를 측정함으로써 목표 분자의 양을 측정할 수 있도록 한다. P_i와 P_i^*는 모두 R_i에 결합할 수 있도록 같은 염기 서열로 디자인 되어있으나, 각각 서로 다른 파장대의 형광을 발산하는 물질이 부착되어있다는 점에서 다르다. 또한 P_i는 모든 i에 대해서 같은 형광 물질을 가지고 있으며 P_i^*도 마찬가지이다.

실제 가중치가 표현되는 메커니즘은 다음과 같다. 가중치-합 연산은 상기 기술한 입력과 그에 대응되는 리포터와 경쟁자 사이에 일어나는 경쟁적 혼성화 반응에 의해 수행된다. 입력 R_i는 P_i 혹은 P_i^*와 만나 각각 입력-리포터 이중가닥 H_i 혹은 입력-경쟁자 이중가닥 H_i^*를 형성한다(그림 2). 이러한 혼성화 반응은 각 R_i에 대해 경쟁적으로 일어나는데, P_i와 P_i^*은 동일한 염기 서열로 구성되어 있고 DNA 분자 간의 혼성화는 해당 DNA의 염기 서열 간의 상보성에 의해 결정 되므로, 상기의 경쟁적인 두 혼성화 반응의 대칭성을 생각할 수 있다. 따라서, 반응 후에 생성되는 H_i와 H_i^*의 양은 반응 초기의 해당 입력의 양과 프로브의 혼합 비율에 의해 다음과 같이 결정된다. 반응 초기 R_i, P_i, P_i^*의 양을 각각 r_i, p_i, p_i^*라 하고 반응 후에 생성되는 H_i와 H_i^*의 양을 각각 h_i, h_i^*라 하자. 충분히 낮은 반응 온도와 입력보다 많은 프로브를 가정할 경우 모든 R_i가 혼성화 반응에 참가한 것으로 생각할 수 있고 두 경쟁적 반응의 대칭성을 고려할 때 반응 후 생성되는 두 이중가닥의 양은 각각 다음과 같다.

$$h_i = r_i \frac{p_i}{p_i + p_i^*},\quad h_i^* = r_i \frac{p_i^*}{p_i + p_i^*}.$$

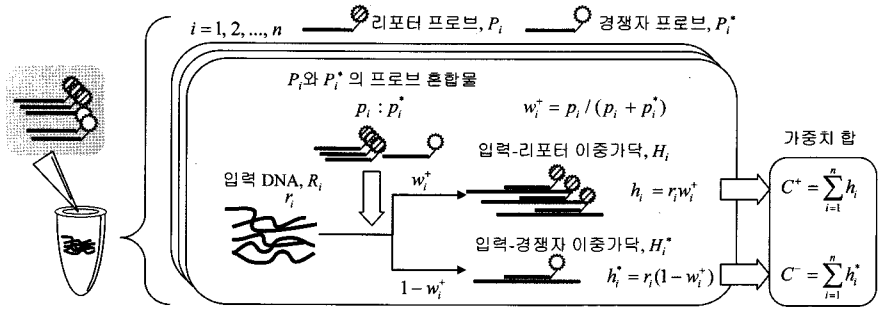


그림 2 DNA 컴퓨팅 기반 생체분자 퍼셉트론을 위한 분자 알고리즘

이 때, 가중치 $w_i^+ = p_i/(p_i + p_i^*)$라 할 때 h_i와 h_i^*는 다음과 같이 다시 쓸 수 있다.

$$h_i = r_i w_i^+, \quad h_i^* = r_i(1 - w_i^+).$$

이러한 경쟁적 혼성화 반응은 모든 입력 요소 R_i에 대해 일어난다. 따라서, 모든 입력-리포터 이중가닥의 총 양과 입력-경쟁자 이중가닥의 총 양은 각각 다음과 같이 가중치-합 연산 결과로 볼 수 있다.

$$C^+ = \sum_{i=1}^{n} h_i = \sum_{i=1}^{n} r_i w_i^+ = \sum_{i=1}^{n} r_i \frac{p_i}{p_i + p_i^*},$$

$$C^- = \sum_{i=1}^{n} h_i^* = \sum_{i=1}^{n} r_i (1 - w_i^+) = \sum_{i=1}^{n} r_i \frac{p_i^*}{p_i + p_i^*}.$$

이 가중치-합 계산 결과 C^+와 C^-는 해당 이중가닥의 총 양을 측정함으로써 얻을 수 있는데, 이는 각각 리포터와 경쟁자에 부착되어 있는 형광 물질에 대응 되는 파장대에서 형광신호의 세기를 측정함으로써 이루어진다[19].

3. 일반화된 혼성화 반응 모델

지금까지 살펴본 가중치의 표현 방법 및 가중치-합 연산을 위한 분자 알고리즘은, 입력-프로브 이중가닥으로 표현되는 계산 결과를 생성하는 한 쌍의 경쟁적인 혼성화 반응이 열역학적으로 대칭이라는 것을 기본 전제로 하고 있다. 이 경우 각 입력 요소 마다 리포터와 경쟁자의 혼합 비율이 직접적으로 바로 가중치를 표현하는 것이다. 이러한 경쟁적 반응의 대칭성은 현재 유전자 발현 패턴의 측정 방법으로서 널리 사용되는 2-색상 마이크로어레이[16] 방법의 바탕을 제공하고 있으며, 여기에 자주 사용되는 두 형광 물질 Cy5와 Cy3를 이용한 프로브들 간의 열역학적 대칭성이 [22]에서 보여진 바 있다. 그러나 혼성화 반응의 열역학적 특성은 반응에 사용된 프로브의 종류, 프로브에 부착된 형광 물질에 따라 달라 질 수 있으며, 다른 형광 물질로 라벨링 되어있거나 분자 비컨[23]과 같은 다른 종류의 프로브를 사용 할 경우 이러한 대칭성의 가정이 성립하지 않을 수도 있다. 여기에서는 경쟁적 혼성화 반응을 비대칭적인 반응으로 확장하여 일반화된 반응 모델을 제안하고, 신뢰성 있는 가중치 표현과 가중치-합 연산을 위한 방법 및 조건을 제시한다.

앞서 언급한 바와 같이 가중치-합 연산을 위한 분자 알고리즘은 다음의 입력 분자 R과 그에 대응되는 프로브인 리포터 P와 경쟁자 P^* 사이에 일어나는 경쟁적인 혼성화 반응을 기본 단위로 하고 있다.

$$R + P \overset{K}{\longleftrightarrow} H$$

$$R + P^* \overset{K^*}{\longleftrightarrow} H^*$$

여기에서 K와 K^*는 각 혼성화 반응의 평형 상수로서, 혼성화 반응 후에 반응 물질과 생성 물질의 양이 더 이상 변화하지 않는 안정 상태에서 반응 물질과 생성 물질의 양 사이의 관계를 나타내는 상수이다[24]. 그러한 안정 상태에서 입력 가닥 R, 리포터 P, 경쟁자 P^*, 그리고 생성된 이중가닥 H와 H^*의 양을 각각 r, p, p^*, h, h^*라고 했을 때, 이들 간에는 다음과 같은 관계가 성립한다.

$$h = Krp, \quad h^* = K^* r p^*$$

그리고 이 두 식으로부터 다음 식을 얻을 수 있다.

$$\frac{h}{h^*} = \frac{K}{K^*}\frac{p}{p^*}$$

여기에서 두 경쟁 반응의 비대칭성을 나타내는 상수로서 $\alpha = K/K^*$를 정의하자. 그러면 반응 초기 리포터와 경쟁자를 포함한 프로브 혼합물의 총 양을 p_T, 프로브 혼합물 가운데 리포터의 비율을 m^+ (경쟁자의 비율은 $1\text{-}m^+$)라고 했을 때, 위 식은 다음과 같이 다시 쓸 수 있다.

$$\frac{h}{h^*} = \alpha \frac{p_T m^+ - h}{p_T(1 - m^+) - h^*}$$

그리고, 입력 R의 초기 양을 r_0라 할 때, 충분히 낮은 반응 온도에서 입력 가닥이 모두 혼성화 되었음을 가정하면 다음 식이 성립한다.

$$r_0 = h + h^*$$

위 두 식을 연립하여 h와 h^*에 대해서 풀어보면 R가 운데 H나 H^*가 되는 비율을 알 수 있으므로, 결과적으로 P_i와 P_i^*의 혼합물에 의해 실제로 표현되는 가중치 $w^+ = h/r_0$를 다음과 같이 얻을 수 있다.

$$w^+ = \frac{h}{r_0} = \frac{2\alpha x m^+}{A + \sqrt{A^2 - 4(\alpha-1)\alpha x m^+}}$$

$$x = \frac{p_T}{r_0},\ A = x + (\alpha-1)(1 + x m^+)$$

두 혼성화 반응이 열역학적으로 대칭일 때(a = 1)는 $w^+ = m^+$가 되어 앞서 기술한 바와 같이 프로브 혼합물에서 리포터의 비율이 바로 실제 원하는 가중치를 표현함을 알 수 있다. 반면 두 반응이 열역학적으로 비대칭일 때($a \neq 1$)는 프로브의 혼합 비율 m^+ 뿐만 아니라, 초기 입력에 대한 프로브의 총 양의 비율 p_T/r_0, 그리고 a에 따른 비선형적인 관계에 의해 가중치가 표현된다(그림 3(a)). 따라서, 가중치 표현을 위한 혼성화 반응이 비대칭적인 특성을 가질 경우, 원하는 가중치를 표현하기 위한 프로브의 혼합 비율 m^+는 이러한 비선형 적인 관계를 고려하여 결정되어야 한다. 그런데 2장에서 언급했듯이 입력 가닥의 완전한 혼성화를 위해서 입력보다 많은 초과량의 프로브를 사용하므로($p_T >> r_0$), 위 결과는 다음과 같이 근사 될 수 있다.

$$\lim_{x\to\infty} w^+ = \lim_{x\to\infty} \frac{2\alpha x m^+}{A + \sqrt{A^2 - 4(\alpha-1)\alpha x m^+}}$$

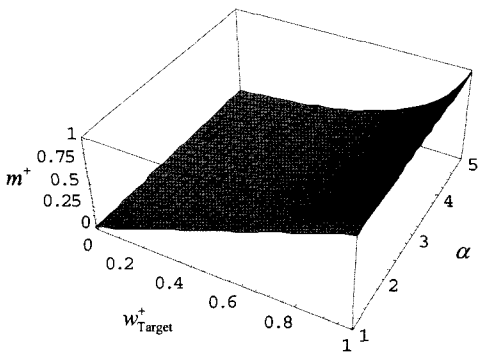


그림 4 입력에 비해 충분히 많은 프로브를 사용할 경우, 다양한 a에 대해 목표 가중치 w^+_{Target}를 표현하기 위한 프로브 혼합 비율 m^+을 결정할 수 있다.

$$= \frac{\alpha m^+}{1 + (\alpha-1)m^+}$$

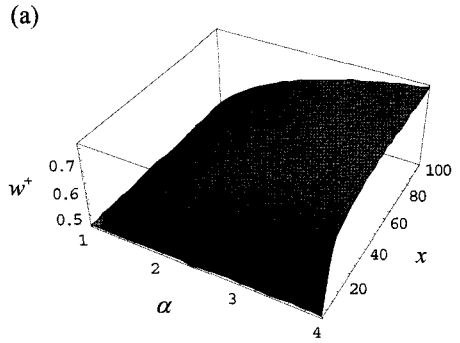
$$\because \lim_{x\to\infty} A/x = \lim_{x\to\infty}\left(1 + (\alpha-1)\left(\frac{1}{x} + m^+\right)\right) = 1 + (\alpha-1)m^+$$

따라서, 초과량의 프로브를 사용할 경우 그림 3(b)에 보이는 바와 같이 프로브 혼합물에 의해 표현되는 가중치는 m^+와 a에 의한 함수가 되어, 목표 가중치 w^+_{Target}를 표현하기 위한 프로브 혼합 비율 m^+를 다음과 같이 결정할 수 있다(그림 4).

$$m^+ = \frac{w^+_{Target}}{\alpha - w^+_{Target}(\alpha-1)}$$

4. 시뮬레이션 결과 및 고찰

앞 장에서 우리는 비대칭적인 열역학적인 특성을 포



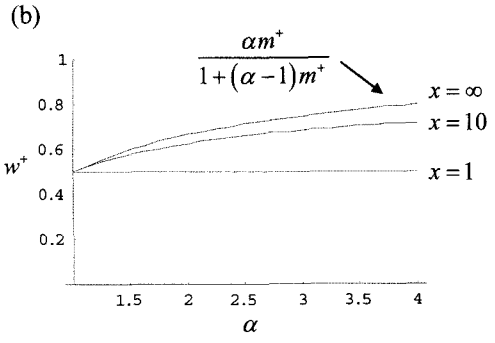


그림 3 프로브 혼합 비율과 그에 따라 표현되는 가중치의 예: m^+ = 0.5. (a) 실제 표현되는 가중치 w^+는, a = 1일 경우 프로브의 혼합 비율 m^+와 일치하지만 그렇지 않을 경우 m^+뿐 아니라, 입력에 대한 프로브의 비율 $x = p_T/r_0$와 a에 따라 비선형적으로 결정된다. (b) 프로브의 양이 입력에 비해 상대적으로 많아지면(x가 커질 때) 이러한 비선형적인 관계는 a와 m^+의 함수가 된다.

함하는 일반화된 경쟁적 혼성화 반응 모델을 제시하였고, 그러한 조건에서 두 반응의 평형 상수 사이의 비율을 고려하여 원하는 가중치를 표현하기 위한 프로브 혼합 비율을 알아보았다. 그러나 앞서 제시된 열역학적 비대칭성을 고려한 가중치 표현 방법은 무한대의 프로브 혼합물을 사용하는 것을 가정하여 계산된 것으로서, 현실적 구현과는 거리가 있으며 실제 구현에 있어서 가중치 표현의 오차를 유발한다. 그러나 입력에 비해 충분히 많은 양의 프로브를 사용할 경우 그러한 오차를 줄일 수 있다. 여기에서는 이렇게 결정되는 프로브 혼합 비율이 실제 가중치를 신뢰성 있게 표현하기 조건을 알아보고, 이전 모델[19]에서 제시된 가중치 표현 방법, 다시 말해 혼성화 반응의 대칭성을 고려하지 않고 프로브 혼합 비율에 의해 직접 표현 되는 가중치 표현 방법과의 성능 차이를 비교 분석한다.

가중치 표현의 성능을 평가하기 위한 지표로서 가중치 표현의 오차 E를 목표 가중치 w^{+}_{Target}와 실제로 경쟁적 혼성화 반응에 따라 표현된 가중치 w^{+}의 차이로 정의하였는데, 이는 다음과 같이 열역학 매개변수 a, 입력에 대한 프로브의 비율 x, 그리고 목표 가중치 w^{+}_{Target}의 함수로 나타내어진다.

$$E\left(\alpha, x, w^{+}_{\mathrm{Target}}\right) = \left| w^{+}_{\mathrm{Target}} - w^{+} \right|$$

이를 바탕으로 다양한 조건하에서 본 논문에서 제안된 새로운 가중치 표현 방법을 기존의 단순히 프로브의 혼합 비율 m^{+}를 가중치로 사용한 경우와 비교하였다. 혼성화 반응을 결정하는 세 개의 변수 a, x, w^{+}_{Target}의 기본 값을 각각 3, 50, 0.5로 두고, 그 가운데 두 개를 고정 시키고 나머지 한 개의 값을 변화시키면서 오차를 비교하였는데, 그 결과가 그림 5에 제시되어있다.

먼저 열역학 매개 변수 a를 변화 시킨 경우(그림 5(a)) a가 커질수록 혼성화 반응의 비대칭성이 심화되어 두 경우 모두 오차가 증가하지만, 비대칭적인 열역학을 고려하여 혼합 비율을 결정했을 경우의 오차가 훨씬 작았다. 가중치는 초기 입력 가닥 가운데 해당 이중가닥으로 변한 비율로 나타내므로 그 최대 값이 1이라는 것을 고려하면, 비대칭성을 고려하지 않은 경우의 오차가 상대적으로 크다는 것을 알 수 있다.

둘째로 프로브 양에 따른 오차의 변화를 살펴보면(그림 5(b)), m^{+}를 w^{+}_{Target}로 직접 사용했을 경우 프로브의 양이 적을 때는 오차가 상대적으로 작은 반면 프로브의 양이 증가함에 따라 오차가 빠르게 증가하였다. 그 이유는 프로브의 양이 작으면 혼성화 반응의 비대칭성에 관계 없이 대부분의 프로브가 혼성화 반응에 참여하게 되

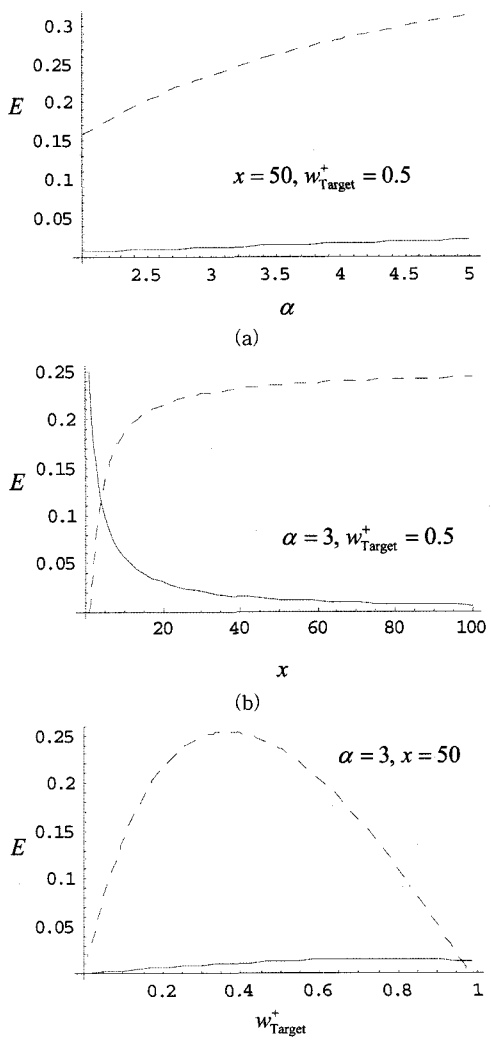


그림 5 혼성화 반응의 비대칭성을 고려한 경우(연속선)와 그렇지 않은 경우(점선)의 오차 비교. 기본 값: x = 50, w^{+}_{Target} = 0.5, a = 3. (a) 열역학 매개 변수 a를 변화 시킨 경우. a가 커질수록 혼성화 반응의 비대칭성이 심화되어 두 경우 모두 오차가 증가하지만, 비대칭적인 열역학을 고려하여 혼합 비율을 결정했을 경우의 오차가 훨씬 작다. (b) 목표 가중치를 직접 혼합비율로 사용했을 경우 프로브의 양이 적을 때는 오차가 작지만 프로브의 양이 증가함에 따라 오차가 증가하며 비대칭성을 고려한 경우의 오차가 더 작아진다. (c) 비대칭성을 고려한 경우 가중치 전 영역에서 고르게 작은 오차를 보이는 반면, 비대칭성을 고려하지 않은 경우에는 중간 영역의 가중치에서 큰 오차가 발생한다.

어 프로브의 혼합 비율이 바로 이중가닥의 생성 비율을 결정하기 때문이다. 반면 비대칭성을 고려한 경우, 프로브의 양이 적을 때는 혼합 비율이 실제 가중치 값과 다르게 정해져 마찬 가지 이유로 오차가 상대적으로 컸지만, 프로브의 양이 늘어남에 따라 결국 오차가 줄어들었는데 이는 앞장에서 제안된 새로운 가중치 표현 방법에서의 초과량의 프로브 사용이라는 가정과 일치한다.

마지막으로 목표 가중치의 변화에 따른 오차를 보면 (그림 5(c)), 비대칭성을 고려한 경우 가중치 전 영역에서 고르게 작은 오차를 보인 반면, 그렇지 않은 경우에는 중간 영역의 가중치에서 큰 오차가 발생하였다. 그 이유는 w^{+}_{Target}를 직접 m^{+}로 사용할 때, 표현하려는 가중치가 0이나 1에 가까울 경우 두 프로브 가운데 한 종류의 양이 입력의 양 보다 작게 되어 혼성화 반응에 의해 모두 고갈 되고 이후 혼성화 반응에서 다른 프로브만이 계속해서 이중가닥을 형성하게 되어 결과적으로 생성되는 이중가닥의 비율 w^{+}가 프로브 혼합 비율 m^{+}에 가까워지기 때문이다.

이상의 결과로부터 혼성화 반응의 비대칭성을 고려해서 프로브의 혼합 비율을 결정했을 경우 더 신뢰성 있는 가중치 표현이 가능하고 모든 가중치를 골고루 잘 표현할 수 있음을 알 수 있다. DNA 기반의 퍼셉트론을 구현하기 위한 프로브의 종류 및 형광 물질이 결정되면, 먼저 일련의 예비 실험을 통해 각 입력 요소에 대응되는 프로브 쌍의 열역학적 특성, 즉, a를 파악하고 이를 바탕으로 그림 4와 같이 목표 가중치를 표현하기 위한 프로브의 혼합 비율을 결정할 수 있다. 그런데, 가중치 표현의 오차는 프로브의 혼합 비율뿐만 아니라 그림 5(b)에서 보여진 바와 같이 각 입력마다 사용하는 프로브의 총 양에 의해서도 영향을 받고, 그 양이 많을수록 오차는 작아진다. 그러나 실제 실험적 구현에 사용 가능한 프로브의 양에는 한계가 있기 때문에 목표 오차 한계를 정하고 그 한계를 만족하기 위해 사용해야 하는 최소한의 프로브의 양을 알아야 할 필요가 있다. 그림 6은 목표 가중치에 대한 함수로서 최대 0.01이하의 오차를 보장하는 프로브의 최소 양을 수치 계산 법을 이용하여 조사한 결과를 보여준다. 결과에서 알 수 있듯이, 평형 상수가 최대 5배 차이가 나더라도 입력보다 100배의 프로브를 넣어준다면 정확한 가중치 표현이 가능하다. DNA 기반 퍼셉트론에서 사용하는 프로브에서 입력에 대한 혼성화를 결정하는 부분의 염기 서열이 동일하고 오직 부착된 형광 물질만 다르다는 점에서 이러한 5배 차이의 평형 상수는 비교적 보수적인 가정이라 생각할 수 있으며 실제 구현에 적용될 때는 더 적은 양의 프로브로도 신뢰성 있는 가중치 표현이 가능할 것으로 예상된다.

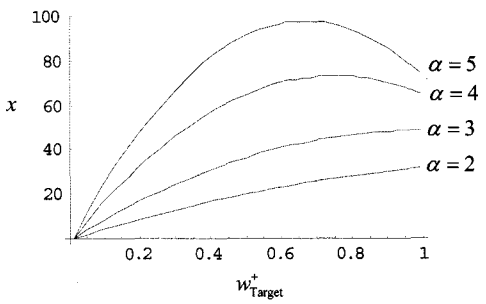


그림 6 0.01 이하의 오차를 보장하기 위한 프로브의 최소 양. 평형 상수가 최대 5배 차이가 나더라도 입력보다 100배의 프로브를 넣어준다면 정확한 가중치 표현이 가능하다.

5. 결 론

본 논문에서는 DNA나 RNA 용액으로 이루어진 분자 패턴의 분류에 적용 가능한 DNA 컴퓨팅 기반의 생체분자 퍼셉트론의 신뢰성 있는 구현을 위한 가중치 표현 방법을 알아보았다. 핵심 연산인 가중치-합 연산을 구현하는 경쟁적 혼성화 반응에서 열역학적인 특성의 비대칭성까지 포함하는 일반화된 경쟁적 혼성화 모델을 제시하였으며 이로부터 목표 가중치를 표현하기 위한 프로브의 구성 방법을 제안하였다. 일련의 시뮬레이션을 통해 반응의 비대칭성을 고려한 경우 목표 가중치의 전 영역에서 안정적인 코딩이 가능하며 단순히 목표 가중치를 바로 혼합 비율로 사용했을 때보다 더 정확한 가중치 표현이 가능함을 입증하였다. 또한 최대 0.01의 가중치 오차를 보장하는데 필요한 최소한의 프로브의 양도 제시하였다.

생체분자 퍼셉트론은 계산적인 관점에서 비교적 단순한 모델임에도 불구하고 DNA 분자들 간의 상보성에 기반한 혼성화 반응을 사용한 계산 모델로서 그 입력으로 이러한 상보성을 가지는 물질은 무엇이든 취할 수 있다는 점에서 폭 넓은 응용 가능성을 가지고 있다. 예를 들어, 유전자의 발현 수준을 보여주는 mRNA나 근래에 새로운 이슈로 떠오르고 있는 micro RNA의 분석에 응용 가능하며, 필요에 따라 입력에 대한 상보적인 결합이 가능한 다양한 프로브를 사용할 수도 있다. 특히 분자 비컨을 이용한 가중치 표현은 분자 알고리즘을 더욱 단순화 시킬 수 있을 것으로 기대 된다. 본 논문에서 제시된 결과를 바탕으로 이러한 생체분자 퍼셉트론의 다양한 응용 및 확장을 위해 효과적인 가중치 표현이 가능할 것으로 기대되며, 경쟁적 혼성화 반응을 바탕으로 하는 수치 정보 표현 방법의 개발에도 도움이 될 것이다.

참 고 문 헌

[1] 이상구, 이광형, 임기형, "DNA 컴퓨팅 기술의 개요 및 응용", 정보과학회지, 제18권, 제8호, pp. 73-77, 2000.

[2] 장병탁, "나노바이오지능 분자컴퓨터: 컴퓨터공학과 바이오공학, 나노기술, 인지뇌과학의 만남", 정보과학회지, 제23권, 제5호, pp. 41-56, 2005.

[3] Păun, G., Rozenberg, G., and Salomaa, A., DNA Computing: New Computing Paradigms, Springer-Verlag, Berlin Heidelberg, 1998.

[4] Adleman, L. M., "Molecular computation of solution to combinatorial problems," Science, Vol.266, No.5187, pp. 1021-1024, 1994.

[5] Lipton, R. J., "DNA Solution of hard computational problems," Science, Vol.268, No.5210, pp. 542-545, 1995.

[6] Braich, R. S., Chelyapov, N., Johnson, C., Rothemund, P. W. K., and Adleman, L., "Solution of a 20-variable 3-SAT problem on a DNA computer," Science, Vol.296, No.5567, pp. 499-502, 2002.

[7] 김은경, 이상용, "해밀톤 경로 문제를 위한 DNA 컴퓨팅에서 코드 최적화", 정보과학회논문지, 제31권, 제4호, pp. 387-393, 2004.

[8] 신수용, 이인희, 장병탁, "ε-다중목적함수 진화 알고리즘을 이용한 DNA 서열 디자인", 정보과학회논문지, 제32권, 제12호, pp. 1217-1228, 2005.

[9] Shin, S.-Y., Lee, I.-H., Kim, D., and Zhang, B.-T., "Multi-objective evolutionary optimzation of DNA sequences for reliable DNA computing," IEEE Transactions on Evolutionary Computation, Vol.9, pp. 143-158, 2005.

[10] Stojanovic, M. N. and Stefanovic, D., "A deoxyribozyme-based molecular automaton," Nature Biotechnology, Vol.21, pp. 1069-1074, 2003.

[11] Benenson, Y., Paz-Elizur, T., Adar, R., Keinan, E., Livneh, Z., and Shapiro, E., "Programmable and autonomous computing machine made of biomolecules," Nature, Vol.414, pp. 430-434, 2001.

[12] Mills Jr., A. P., Yurke, B., and Platzman, P. M., "Article for analog vector algebra computation," BioSystems, Vol.52, No.1-3, pp. 175-180, 1999.

[13] Sakakibara, Y. and Suyama, A., "Intelligent DNA chips: logical operation of gene expression profiles on DNA computers," Genome Informatics, Vol.11, pp. 33-42, 2000.

[14] Mills Jr., A. P., "Gene expression profiling diagnosis through DNA molecular computation," Trends in Biotechnology, Vol.20, No.4, pp. 137- 140, 2002.

[15] Benenson, Y., Gil, B., Ben-Dor, U., Adar, R., and Shapiro, E., "An autonomous molecular computer for logical control of gene expression," Nature, Vol.429, No.6990, pp. 423-429, 2004.

[16] Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J., "High density synthetic oligonucleotide arrays," Nature Genetics, Vol.21, pp. 20-24, 1999.

[17] Oliver, J. S., "Matrix multiplication with DNA," Journal of Molecular Evolution, Vol.45, No.2, pp. 161-167, 1997.

[18] Mills, Jr., A. P., Turberfield, M., Turberfield, A. J., Yurke, B., and Platzman, P. M., "Experimental aspects of DNA neural network computation," Soft Computing, Vol.5, pp. 10-18, 2001.

[19] Lim, H.-W., Lee, S. H., Yang, K. A., Lee, J. Y., Yoo, S.-I., Park, T. H., and Zhang, B.-T., "In vitro molecular pattern classification using DNA- based weighted sum computation," IEEE Transactions on Nanobioscience, (submitted for publication).

[20] Rosenblatt, F., "The perceptron: a probabilistic model for information storage and organization in the brain," Psychological Review, Vol.65, No.6, pp. 386-408, 1958.

[21] McCulloch, W. S. and Pitts, W., "A logical calculus of the ideas immanent in nervous activity," Bulletin of Mathematical. Biology, Vol.5, pp. 115-133, 1943.

[22] Wang, Y., Wang, X., and Goush, S.-W., "Conditions to ensure competitive hybridization in two- color microarray: a theoretical and experimental analysis," BioTechniques, Vol.32, No.6, pp. 1342- 1346, 2002.

[23] Tyagi, S. and Kramer, F. R., "Molecular beacons: probes that fluorescence upon hybridization," Nature Biotechnology, Vol.14, No.3, pp. 303-308, 1996.

[24] Rose, J. A., Deaton, R. J., and Suyama, A., "Statistical thermodynamic analysis and design of DNA-based computers," Natural Computing, Vol.3, No.4, pp. 443-459, 2004.

임 희 웅

1999년 2월 서울대학교 전산학과 학사 2001년 2월 서울대학교 전기컴퓨터공학부 석사. 2007년 2월 서울대학교 전기컴퓨터공학부 박사. 관심분야는 Molecular/DNA Computation, 생물정보학, 기계학습

유 석 인

정보과학회논문지 : 소프트웨어 및 응용 제 34 권 제 11 호 참조

장 병 탁

1986년 서울대학교 컴퓨터공학 학사. 1988년 서울대학교 컴퓨터공학 석사. 1992년 독일 Bonn대학교 컴퓨터공학 박사. 1992년~1995년 독일국립정보기술연구소(GMD) 연구원. 1995년~1997년 건국대학교 컴퓨터공학과 조교수. 1997년~현재 서울대학교 컴퓨터공학부 교수, 인지과학, 뇌과학, 생물정보학 협동과정 겸임. 관심분야는 Biointelligence, Probabilistic Model of Learning and Evolution, Molecular/DNA Computation