바이오데이터 분석을 위한 기계학습 기술

숭실대학교 ▌황규백*
서울대학교 ▌정제균 · 남진우** · 김병희** · 이제근** · 장병탁***

1. 생명과학 연구와 기계학습

최근 생명공학의 급속한 발전으로 대규모의 바이오데이터가 생성됨에 따라 이를 분석하는 컴퓨터 기술의 중요성이 더욱 증가하고 있다. 컴퓨터를 사용하여 생명현상을 연구하는 컴퓨터생물학(computational biology)과 바이오데이터를 처리하기 위한 정보기술을 다루는 생물정보학(bioinformatics) 연구가 더욱 활성화되고 있을 뿐만 아니라 정보기술의 발전과 함께 생명과학의 새로운 패러다임이 등장하고 있다. 개개의 유전자나 단백질을 다루던 종래의 연구에서 한 걸음 더 나아가 유전체(genome) 혹은 단백체(proteome) 전체를 다루는 이른바 Omics 연구가 가능하게 되었으며, 세포내의 전체 분자들의 상호작용망을 분석하려는 이른바 네크워크 생물학(network biology) 또는 시스템생물학(systems biology) 연구 분야가 새로이 등장하고 있다. **최근 들어, 기계학습은 이러한 생명과학 및 의약학 연구를 위한 바이오데이터 마이닝 및 모델링을 위한 핵심 기반기술로 자리매김 되었다.**

본고는 생명과학 분야의 다양한 문제 해결을 위하여 기계학습 기술이 적용된 사례들을 살펴보는 것을 주목적으로 한다. 2절에서는 감독학습(supervised learning) 기반의 바이오데이터 분석을 다룬다. 특히, 최신의 기계학습 기법인 베이지안망(Bayesian network)과 커널 머신(kernel machine)이 바이오데이터 분석에 적용된 사례를 기술하고, 생명과학 분야의 주요문제 중 하나인 모티프 예측에 감독학습 기법이 적용된 사례를 상세히 설명한다. 3절에서는 무감독학습(unsupervised learning) 기반의 바이오데이터 분석을 다룬다. 다양한 군집화(clustering) 기법, 확률그래프모델 및 잠재변수모델(latent variable model)이 바이오데이터 분석에 적용된 사례들을 기술한다. 4절에서는 다양한 기계학습 기법을 적용한 질병진단 사례들을 상세히 설명하며, 5절에서 결론을 맺는다.

본론에 들어가기에 앞서 여기서 생명과학 연구에서 다루는 주요 바이오데이터의 특성과 현안 문제들을 간략히 살펴보기로 한다(그림 1). 컴퓨터 자료구조 관점에서 볼 때, **바이오데이터는 크게 다섯 가지 종류로 구분해 볼 수 있다. 이것은 (1) 1차원 배열형태의 서열 데이터, (2) 다차원 배열 형태의 구조 데이터, (3) 매트릭스 형태의 발현 데이터, (4) 네트워크 형태의 상호작용 데이터, (5) 문서 형태의 텍스트 데이터이다.**

서열 데이터의 예로는 DNA 및 RNA 등의 유전체 데이터, EST(expressed sequence tag) 서열, SNP(single nucleotide polymorphism) 데이터 등이 있으며, 이들은 네개의 문자 A, G, C, T(또는 U)로 구성된 스트링 형태

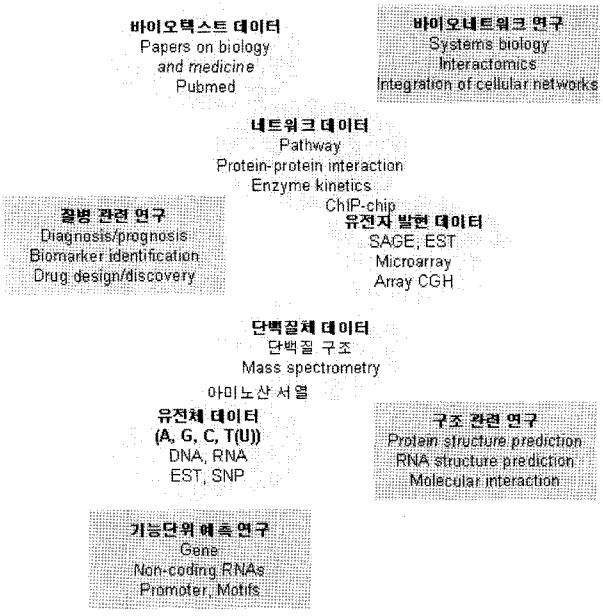

그림 1 다양한 바이오데이터와 생명과학 연구주제

† 본 논문은 2006년도 과학기술부의 재원으로 한국과학재단의 국가지정연구실사업의 연구결과로 수행되었습니다.
(No. M10400000349-06J0000-34910)
* 정회원
** 학생회원
*** 종신회원

이다. 구조 데이터의 예로는 단백질 3차 구조 데이터, 질량분석기(mass spectrometry) 데이터 등이 있으며, 이는 각각 3차원 상의 위치정보와 특정 효소에 의해 조각난 단백질의 질량을 표현한다. 발현(expression) 데이터는 SAGE(serial analysis of gene expression) 데이터나 EST 데이터를 비롯해 수천에서 수만에 이르는 유전자의 발현량을 한꺼번에 측정할 수 있는 마이크로어레이(microarray) 데이터가 있다. 최근에는 마이크로어레이 기술로 개개의 유전자의 발현량 뿐 아니라 유전체의 특정지역의 양(copy number)의 세밀한 측정이나 특정 단백질의 결합 부위를 대규모로 검색할 수 있는 array CGH(comparative genomic hybridization) 및 ChIP(chromatin immuno-precipitation)-chip 데이터 또한 생산되고 있다. 네트워크 데이터는 분자들 간의 반응 또는 상호작용을 나타내는 pathway 데이터, protein interaction 데이터 및 동적 signal pathway 분석을 위한 enzyme kinetics 데이터 등이 있다. 텍스트 데이터는 논문 등과 같은 문서 형태로 정리된 생물학 정보를 담은 정보로서 PubMed와 같은 데이터베이스가 그 대표적인 예이다.

기계학습 기술이 적용된 생명과학 문제는 분자생물학적인 기초연구로부터 질병진단 및 신약개발의 응용에 이르기까지 매우 다양하다. DNA 서열정보에 기반하여 유전체 내에서 유전자를 인식하는 유전자 예측 문제는 이미 오래전부터 신경망이나 은닉마코프모델과 같은 기계학습 기법이 적용된 고전적인 문제 중의 하나이다. 그 외에 유전체 분석과 관련하여 프로모터 예측, 유전자 전사조절 인자(transcription factor) 결합 위치 분석, miRNA 예측 등의 문제에도 기계학습 알고리즘이 활용되고 있다. 단백체학 관련하여서는 단백질 구조 예측, 단백질 기능 예측, 단백질 간의 결합, 혹은 단백질과 다른 물질과의 결합을 예측하는 문제 등이 중요하다. 네트워크 관점에서는 유전자나 단백질의 관계를 밝히는 문제, 유전자 사이의 조절관계망이나 단백질 상호작용망 등을 구성하기 위해 기계학습이 사용된다. 대량의 유전체나 단백체 데이터에 기반하여 질병진단을 하는 문제의 경우 역시 기계학습이 중요한 역할을 한다. 단순한 질병진단 뿐 아니라 질병발생과 관련된 요인들이 무엇인지 판별하는 것 역시 이 분야의 주요과제이다. 질병발생과 관련된 요인의 분석은 더 나아가 신약개발과도 관련이 되어 있다.

2. 감독학습 기반의 바이오데이터 분석

2.1 베이지안망

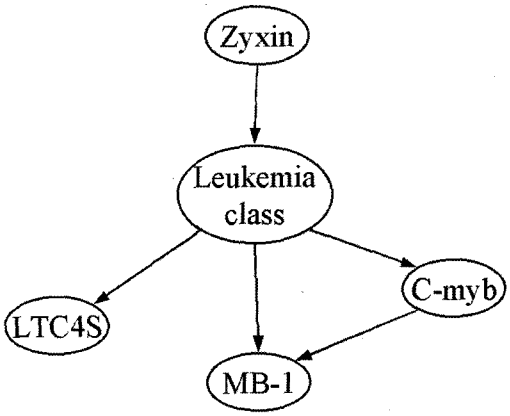

그림 2 급성백혈병 분류를 위한 베이지안망[1]. 백혈병의 종류를 나타내는 Leukemia class 노드와 4개의 유전자 노드로 구성되어 있다.

변수들 사이의 조건부독립성(conditional independence)에 기반하여 결합확률분포를 효율적으로 표현하는 베이지안망은 DAG(directed acyclic graph) 구조로 되어 있으며 노드는 변수를, 간선은 변수들 사이의 의존관계(dependency)를 나타낸다. 베이지안망이 주어진 경우, n개의 변수 $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$의 결합확률분포는 아래와 같이 표현된다.

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i|\mathbf{Pa}(X_i)) \quad (1)$$

위 식에서 $\mathbf{Pa}(X_i)$는 그래프에서 변수 X_i의 부모노드 집합을 가리킨다.

분류문제를 해결하기 위한 베이지안망은 그래프 구조에 주어지는 제한의 정도에 따라 나이브 베이즈 분류기(naive Bayes classifier), TAN(tree-augmented naive Bayes) 분류기 및 일반적인 베이지안망 분류기로 나누어 볼 수 있다. 그래프에 주어지는 제한이 강할수록 학습은 용이하나 그 표현력은 떨어지게 된다.

베이지안망 분류기는 다양한 생명과학 문제 중 마이크로어레이 데이터를 통한 질병진단에 주로 적용되었다. 급성백혈병(acute leukemia)의 종류를 구분하는데 일반 베이지안망 분류기(그림 2)가 사용되었으며[1], 급성백혈병 및 결장암 마이크로어레이 데이터 분석을 위해 베이지안망 분류기의 앙상블을 사용한 예가 보고되었다[2]. 또한 잡음이 심하고 희박한(sparse) 마이크로어레이 데이터를 다룰 때 일반화 성능을 향상시키기 위한 BMA(Bayesian model averaging) 기법도 개발된 바 있다[3].

마이크로어레이 데이터 분석 외에 서열 및 구조정보를 가지고 단백질을 분류하는 문제에도 베이지안망 분류기가 적용되었다. 기존의 은닉마코프모델(hidden

Markov model)보다 뛰어난 성능을 보이는 베이지안망 분류기가 고안되었으며[4], 전사조절 인자(transcription factor) 결합 위치를 예측하는데 베이지안망이 적용되었다[5]. 또한 베이지안망을 이용하여 단백질, 유전자에 관련된 여러 정보들을 결합하여 유전자 기능 예측 성능을 향상시킬 수 있음도 보고된 바 있다[6].

2.2 커널 기법

기계학습에서 커널 기법은 데이터에 커널 함수를 적용하여 고차원의 공간으로 사상(mapping)시킨 후 패턴 분석을 하는 기법이다. 분류 문제를 위한 응용의 관점에서 보면 원래의 데이터를 선형적인 분류함수 적용이 가능한 고차원 공간으로 사상(mapping)시키는 것을 의미한다(그림 3). 즉, 커널 기법을 통해 사상된 공간에서의 선형 분류가 기존 공간에서의 비선형 분류와 동등하게 된다. 감독학습 기반의 대표적인 커널 기법은 support vector machine(SVM)이다. SVM 기법에서는 커널을 이용하여 고차원으로 사상된 공간에서 데이터를 선형적으로 분류하기 위한 maximal-margin hyperplane을 학습한다.

커널 기법은 표 1과 같이 마이크로어레이 데이터 기반의 조직 분류 및 질병진단을 비롯하여, 단백질 분류 등 다양한 생명과학의 문제에 적용되었다.

SVM은 최근 주목 받고 있는 microRNA(miRNA)의 목표유전자 예측에도 적용되었다. miRNA는 유전자의 3' UTR(untranslated region) 영역에 결합하여 유전자의 기능을 억제하는 역할을 수행한다. 특정 miRNA가 mRNA(messenger RNA, 일반적인 유전자)와 결합했을 때의 2차 구조 정보에 기반하여, 특정 miRNA가 특정 유전자의 발현을 억제할 수 있는지 여부를 커널 기법으로 예측한 것은 감독학습 기법이 생명과학의 주요 문제 해결에 성공적으로 적용된 대표적인 예이다[12].

자질선정이 가능한 p-SVM(potential SVM) 방법은 마이크로어레이 데이터를 이용하여 암 분류에 중요한 영향을 미치는 유전자를 찾아내는 연구 등에 응용되기도 하였다[13].

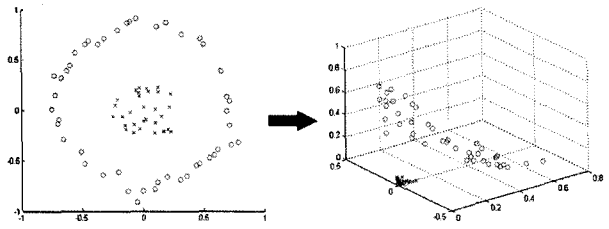

그림 3 커널을 이용한 고차원으로의 데이터 사상

표 1 생명과학 문제에 적용된 커널 기법의 예

커널 종류	생물학 응용 예	참고 문헌
Polynomial Kernel	Tissue classification	[7]
RBF Kernel	Tissue classification	[7]
Spectrum Kernel	Protein family classification	[8]
Tree Kernel	Phylogeny analysis & gene function prediction	[9]
Local Alignment Kernel	Protein homology detection	[10]
Graph Kernel	Prediction of chemical compound properties	[11]

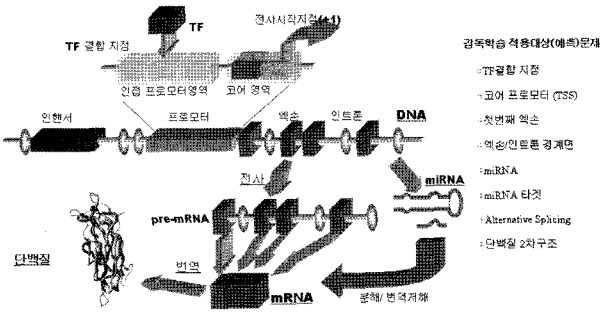

그림 4 miRNA 조절을 반영한 현대 분자생물학의 central dogma. 유전정보가 단백질로 발현되는 과정에 대한 연구 대부분에 감독학습 기법이 주요 도구로 사용되고 있다.

2.3 인공신경망 기반 프로모터 예측

감독학습 기법은 현대 분자생물학의 central dogma인 유전자가 전사(transcription)되어 단백질로 번역(translation)되는 기작에 관련된 다양한 예측문제 해결에 핵심도구로 사용되고 있다(그림 4). 이 절에서는 **감독학습의 대표적 기법인 인공신경망(artificial neural network)을 적용하여, DNA 서열상에서의 유전자 발현 조절 단계의 핵심 영역인 '프로모터'를 예측한 사례(PromSearch[14], 그림 5를 통해 감독학습 기법을 이용한 예측문제 해결 과정을 살펴본다.**

PromSearch는 다음과 같은 과정을 통해 프로모터 영역을 판별하고 전사시작 지점을 예측한다.

1) 문제정의/모델 설정: DNA서열상에서 프로모터와 비(非)프로모터를 구분하는 이진 분류문제를 정의. 공통적으로 존재하는 모티프(예: TATA 박스) 신호와 모티프 집중 분포지역의 통계적 특성을 반영하여, 프로모터 주변 300bp 영역에 대한 모델을 설정한다.

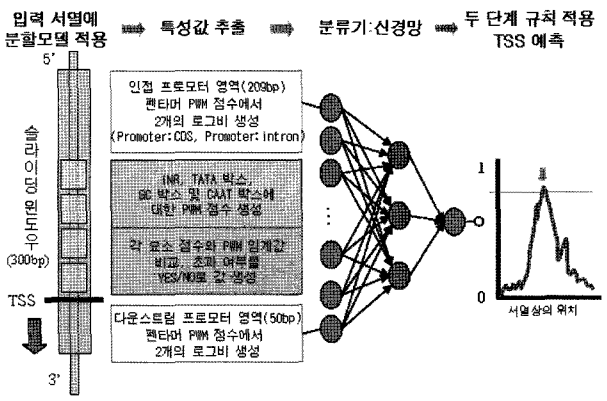

그림 5 감독학습의 바이오데이터 분석 적용사례: 유전자 발현을 조절하는 프로모터 영역 예측에 '인공신경망'이 핵심 도구로 사용되었다[14].

2) 설정한 모델의 요소별 자질(feature) 추출 : PWM (position-weight matrix)을 이용하여 서열의 특성을 수치화한다.
3) 인공신경망 학습 : 수치화된 특성값을 이용해 인공신경망을 구성한다. 프로모터서열 DB의 데이터에서 positive 데이터를, 프로모터 이외의 유전자 영역(엑손, 인트론)의 서열에서 negative 데이터를 구성하여 인공신경망을 학습한다.
4) 적용 : 미지의 DNA 서열에 대해 300bp의 슬라이딩 윈도우를 이동시켜가며, (1), (2)의 과정을 통해 자질값을 추출하고, (3)에서 학습한 인공신경망에 입력하여 프로모터 여부를 판별한다.
5) 인공신경망이 프로모터로 판별하는 경우, 전사시작지점(transcription start site, TSS)의 위치를 결정한다.

3. 무감독학습 기반의 바이오데이터 분석

3.1 군집화

바이오데이터 분석에 있어서 무감독학습 기반의 군집화 알고리즘은 주로 마이크로어레이 데이터에서 발현 패턴이 유사한 유전자군을 탐색하는 데 적용되었다. 함께 발현(co-expression)되는 유전자들은 생물학적으로 비슷한 기능을 하거나 함께 조절(co-regulation)될 가능성이 높으며, 이는 유전자 망 구성 및 유전자와 단백질의 기능 예측을 위한 기초적인 정보가 된다.

군집화 기법은 크게 분할 군집화(partition clustering), 계층적 군집화(hierarchical clustering), 주성분분석(principal component analysis, PCA), 혼합모델(mixture model), 공군집화(co-clustering)로 구분되며 각 기법의 개념과 응용의 예는 다음과 같다.

1) 분할 군집화 : 분할 군집화에서는 주어진 데이터를 미리 정해진 개수의 군집으로 분할하여 군집화를 하게 된다. 그 대표적인 알고리즘인 k-평균(k-means)은 같은 군집 내의 데이터에 대해 제곱합(sum of square)이 최소화되도록 분할한다. 퍼지 c-means는 k-평균과 유사하나, 하나의 유전자를 여러 군집에 할당할 수 있는 특징을 가지고 있다. 자기조직화지도(self-organizing map, SOM)는 데이터를 2차원 격자 상에 사상시킨다. 분할 군집화 기법은 마이크로어레이 데이터 분석 및 조절인자의 결합 위치 정의에 적용되었다[15-18].

2) 계층적 군집화 : 계층적 군집화는 덴드로그램(dendrogram)이라는 트리 형태로 데이터를 분할한다(그림 6). 트리를 생성하는 방법에 따라 병합식(agglomerative)과 분할식(divisive)으로 나눌 수 있다. 계층적 군집화는 마이크로어레이 데이터 및 서열 분석에 활용되었으며[19,20], 최근에는 커널 기법을 도입하여 고차원 자질(high-order feature)을 효과적으로 다루려는 시도가 있었다[21].

3) 주성분분석 : 주성분분석은 고차원 데이터의 구조를 밝히거나 차원을 낮추는데 이용되는 다변량 통계분석 방법이다. 최근 마이크로어레이 데이터에 빈번한 이상치를 효과적으로 제거하기 위한 견고한 주성분분석(robust PCA) 기법이 제안되었다[22].

4) 혼합모델 : 혼합모델은 파라미터를 가지는 다수의 함수들의 합으로 확률밀도함수를 모델링하여 군집화를 실현한다. 혼합모델은 전통적인 통계적 분석 기법에 해당하며, 최근 마이크로어레이, SNP 및 전사인자 결합위치 등의 분석에 다양하게 적용되고 있다[23,24].

5) 공군집화 : 공군집화는 주어진 데이터 행렬의 행과 열을 동시에 군집화하는 기법으로 biclustering이라고도 불린다. 마이크로어레이 데이터 분석에서는 전체 표본이 아닌 부분 표본에 대해서만 유사한 양상을 보이는 유전자들의 군집화에 활용된다[25].

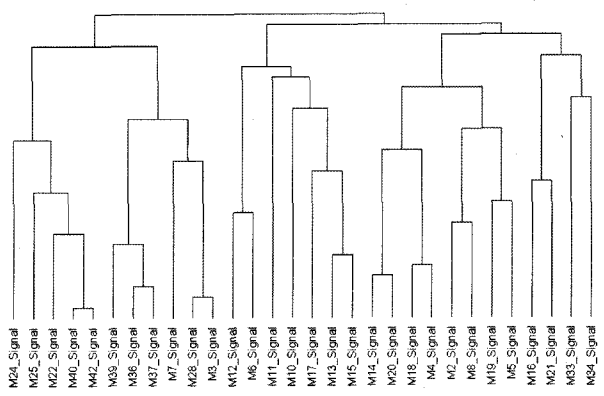

그림 6 계층적 군집화의 결과 덴드로그램

표 2 군집화 알고리즘을 이용한 바이오데이터 분석 사례

구분	알고리즘	적용 사례
분할 군집화	*k*-means	Microarray analysis [15]
	Fuzzy c-means	Microarray analysis [16]
	SOM	Microarray analysis [17] Sequence analysis [18]
계층적 군집화	Hierarchical clustering	Microarray analysis [19] Sequence analysis [20]
	Kernel hierarchical clustering	Microarray analysis [21]
성분 분석	PCA	Microarray analysis [22]
혼합 모델	Mixture model	Microarray analysis [23] Sequence analysis [24]
공 군집화	Biclustering	Microarray analysis [25]

표 2에 위에서 기술한 다양한 군집화 기법이 바이오데이터에 적용된 사례를 정리하였다.

3.2 확률그래프모델 및 베이지안망

베이지안망을 비롯한 확률그래프모델은 결합확률분포를 표현하므로 감독학습뿐 아니라 무감독학습에도 적용될 수 있다. 확률그래프모델은 바이오데이터에서 바이오네트워크를 추론하는데 활용될 수 있으며, 최근 활발히 연구되고 있는 시스템생물학(systems biology) 분야에 적합한 방법론으로 대두되고 있다.

효모의 시간상(temporal)의 유전자 발현양상을 기록한 마이크로어레이 데이터에서 유전자 간의 상호작용을 모델링하는 베이지안망을 학습한 예가 있으며[26], 동적 베이지안망(dynamic Bayesian network)을 이용하여 유전자망을 학습한 예도 있다[27]. 이러한 연구는 시계열(time-series)분석을 통해서 단순히 유전자 간의 연관성을 보는 것이 아니라, 조절망(regulatory pathway) 상에서 유전자들이 어떻게 상호 조절을 하는지 전체적으로 파악할 수 있는 직관적인 모델을 제공한다는 장점이 있다.

유전자 발현 데이터에 약물의 반응 자료를 결합하여 약물 반응과 유전자 발현 사이의 관계를 베이지안망으로 학습한 연구결과도 제시되었다[28]. 구체적으로 베이지안망의 구조학습 뿐 아니라 확률적 추론을 적용하여 유전자와 약물 반응 사이의 관계를 계량화하였다. 이 연구에서는 종양 세포에서 유전자 발현과 약물의 영향력 간의 관계를 학습한 베이지안망(그림 7)을 통해 생물학적으로 의미있는 새로운 관계를 밝혀냈다.

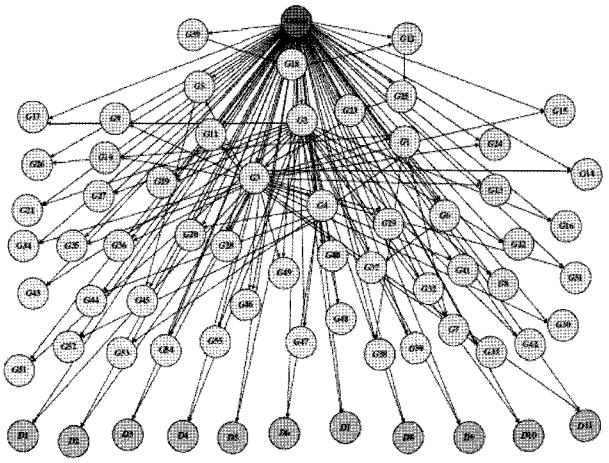

그림 7 유전자발현-약물반응 관계를 학습한 베이지안망[28]

또한 보다 정확한 유전자망의 학습을 위해 필요한 실험을 제안할 수 있는 능동학습(active learning)기법이 개발되었으며[29], 다양한 확률그래프모델의 유전자망 학습에 있어서의 성능도 비교분석되었다[30].

확률그래프모델은 다양한 특성의 정보를 자연스럽게 결합할 수 있는 장점이 있으며, 학습 결과의 신뢰도 향상을 가능하게 한다. 예를 들어, 단백질-단백질 상호작용 데이터를 추가하여, 유전자 발현패턴 분석의 신뢰도를 향상시키고 보다 정밀한 유전자망을 추론하는 기법이 소개되었다[31,32].

3.3 잠재변수모델

잠재변수(latent variable)는 데이터의 차원 축소, 미관측 데이터/요소에 대한 설명, 개념적인 변수의 표현 등을 위해 도입되며, 계산의 간소화, 분석결과의 가시화 등의 효과를 준다. 바이오데이터에는 잡음이 많으며, 데이터로 측정할 수 없는 요인들이 존재하는 경우도 많기 때문에, 잠재변수를 포함하여 데이터를 분석하는 잠재변수모델(latent variable model, LVM)은 바이오데이터 분석에도 많이 활용되고 있다.

최근 잠재변수모델로 공군집화의 특성을 포함하는 공군집화 LVM이 제안되었다[33]. 공군집화 LVM은 잠재변수를 도입하여 두 대상의 확률적 상관관계를 파악하는 모델이며, 줄기세포의 조절 기작을 밝히는데 적용되었다. 줄기세포 연구에서는 특정 세포로 분화하는 조절 기작의 이해가 주요 관심사이다. 기존의 대부분의 연구는 단지 몇 개의 조절자(regulator)에 집중하여 거시적인 결과를 얻기 어려웠던 반면, 이 연구에서는 줄기세포 관련 마이크로어레이와 유전체 서열 데이터를 통합한 공군집화를 통해 세포특이적인 조절인자와 이들이 결합하는 목표유전자의 관계를 대규모로 밝혀

냈다(그림 8, 9). 제시된 관계성은 대부분 기존 문헌의 결과와 일치하며, 유전자 조절 과정에 대한 흥미로운 새로운 가설을 제시하고 있다. 이 기법은 줄기세포 뿐만 아니라 다른 세포 특이적 조절인자 연구에도 유용한 결과를 제공할 것으로 기대되고 있다.

이외에도 유전자 발현 프로파일에 기반하여 베이지안망을 이용한 대규모 유전자 조절망을 구축하고자 한 연구[34]에서는, 유사 유전자군을 대표하는 잠재변수를 도입하여 계층적 구조의 유전자망을 구축하였다.

SOM과 확률적 LVM의 특성을 결합한 모델인 SOLL[35]은, 시간에 따른 발현 프로파일의 변화를 반영하여 공통의 발현 패턴을 가지는 유전자 발견과 패턴의 가시화를 가능하게 하였다.

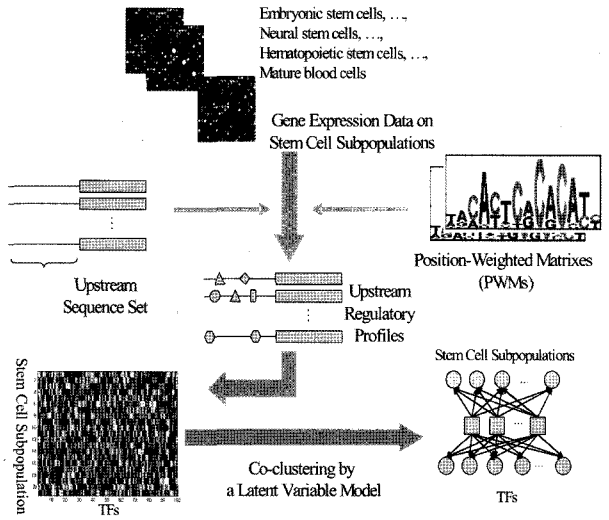

그림 8 공군집화 LVM을 이용한 줄기세포군과 조절인자간의 관계 규명 개념도[33]

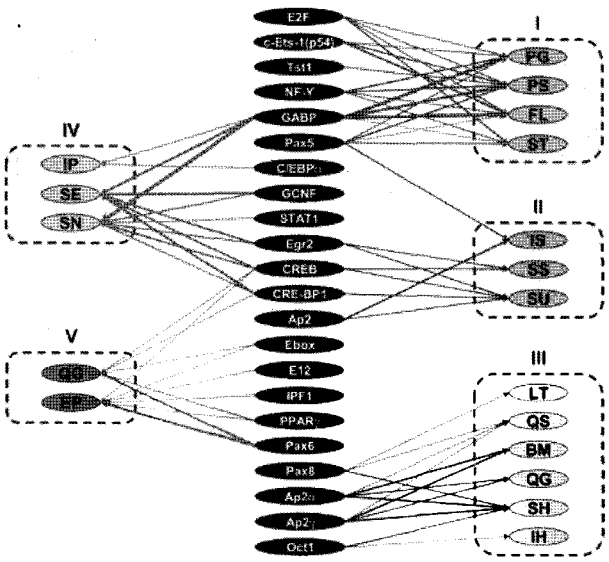

그림 9 공군집화 LVM으로 찾아낸 줄기세포 관련 조절모듈[33]

4. 질병진단 및 치료를 위한 기계학습 기술

4.1 압타머(aptamer) 분석을 이용한 질병진단

혈액검사는 신체의 상태 파악을 위한 기초자료로 많이 활용되고 있다. 질병의 여부에 따라 말초혈액(peripheral blood)내의 유전자 발현양상에 변화가 발생하기도 하며[36], 혈액 내에 특정 질병의 지표단백질들이 존재한다는 사실도 알려져 있다[37]. 이와 같이 혈액 내 단백질 발현양상의 변화는 다양한 질병진단을 위한 주요 정보를 담고 있다.

특정 단백질을 동정(同定, identification)하기 위한 기법 중 하나로, 압타머(aptamer)를 이용하는 기술이 있다. **압타머란 단일염기서열 상태의 DNA나 RNA로, 항원-항체 반응과 같이 목표물질에 대해 특별한 친화력과 특이성을 보이는 생체정보 감지소재이다.** 제노프라(주)[1]에서는 압타머를 이용한 단백질 마이크로어레이인 압타머칩을 개발하였으며, 혈액 내 수천여 개의 단백질 분포의 변화를 살펴봄으로써 질병 진단을 가능하게 하였다.

압타머칩을 이용해 간암, 심혈관질환 등의 진단이 시도되고 있으며, 여기에 다양한 기계학습 기법이 핵심 도구로 활용되고 있다. 심혈관질환의 단계 구분에 인공신경망, 결정트리, SVM 및 베이지안망의 네 가지 감독학습 기법이 적용되어 높은 분류 성능을 얻었으며(그림 10)[38,39], 특정 질병의 지표(marker) 압타머/단백질 탐색에 기계학습의 주요 연구 분야인 자질 선정(feature selection) 기법이 적용된 바 있다[40].

4.2 Small RNA를 이용한 질병진단

Small RNA는 길이가 약 18nt(nucleotide)에서 30nt 정도 되는 작은 non-coding RNA 집단을 말한다. 최근 non-coding RNA인 miRNA, siRNA의 유전자 조절기능이

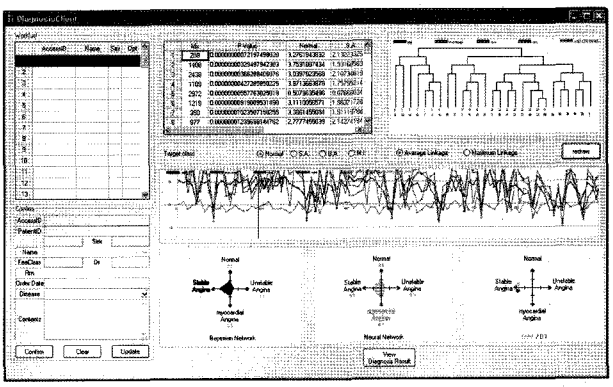

그림 10 AptaCDSS-압타머칩을 이용한 심혈관질환 질환단계 예측 및 진단의사결정지원시스템

1) http://www.genoprot.com

밝혀지면서, siRNA나 miRNA를 이용한 RNA interference (RNAi) 기술이 유전자 knock-down 시스템, 유전자치료 등에 활발히 응용되고 있다[41,42]. **이러한 응용분야에 다양한 기계학습알고리즘이 적용되고 있다. Genetic programming을 이용한 높은 RNAi 효용성을 가지는 siRNA 설계 알고리즘은 그 대표적인 예이다[43].**

최근, 전사후 조절인자(post-transcriptional regulator)로 주목받고 있는 miRNA는 암과 관련된 유전자를 직접 조절함으로써 암의 발생과 직접적으로 연관되어 있으며, miRNA의 발현양상은 기존의 유전자(mRNA) 발현양상을 이용한 암의 분류보다 더 뛰어난 분류자로 사용될 수 있음이 보고되었다[44-49]. **질병진단과 관련이 깊은 miRNA-유전자 모듈을 찾기 위한 노력으로 베이지안망과 같은 확률그래프모델이나[50], 공군집화를 위한 유전알고리즘[51]이 소개되고 있다.**

4.3 인유두종바이러스(HPV)에 기반한 자궁경부암 진단 기술

인유두종바이러스(human papillomavirus, HPV)는 약 8kb의 환상(環狀)의 이중나선 DNA 바이러스로 여성의 자궁경부암을 유발하며 여러 가지 악성종양과 밀접한 관계가 있는 것으로 알려져 있다. HPV는 지금까지 85종에 달하는 유전형(genotype)의 염기서열이 완전히 밝혀져 있으며 120여 개의 새로운 HPV유전형 구조가 부분적으로 보고되고 있다[52]. HPV의 분류는 DNA 염기서열의 유사성에 따르고 있는데 E6, E7, L1 ORF(open reading frame)의 염기서열(그림 11)이 기존에 보고된 염기서열과 10% 이상의 차이를 보이면 새로운 유전형으로, 90-98% 유사성을 보이는 경우는 아형(subtype)으로, 98% 이상의 유사성에 대해서는 동일 유전형 내 변체(variant)로 정의하고 있다.

특히 자궁경부와 관련된 HPV는 악성종양 유발 가능성에 따라 고위험군(high-risk type)과 저위험군(low-risk type)으로 나뉜다[53]. 예를 들어, HPV 16, 18, 31과 같은 고위험 HPV에 감염될 경우 악성종양으로 진행될 가능성이 높다[54]. 따라서, 감염된 HPV의 형(type)을 파악하는 것이 환자의 처방 및 예후에 매우 중요하다. 기존에는 이러한 HPV의 위험군 분류를 수행할 때 생물학자가 직접 수많은 문헌자료로부터 조사하는 수작업이 필요했다. 하지만 **최근에는 텍스트마이닝 기술의 발달에 따라 컴퓨터를 이용한 자동분류가 가능해졌다. 그 예로 결정트리를 이용한 텍스트마이닝 기법이 HPV의 자동 분류에 적용된 사례가 있다[55].** 이 시스템은 먼저 대량의 관련 문서에서 자동적으로 위험군을 분류한 다음 해당 분야의 전문가가 검증하는 절차를 거친다는 점에서 분류 작업의 효율을 높일 수 있는 장점이 있다.

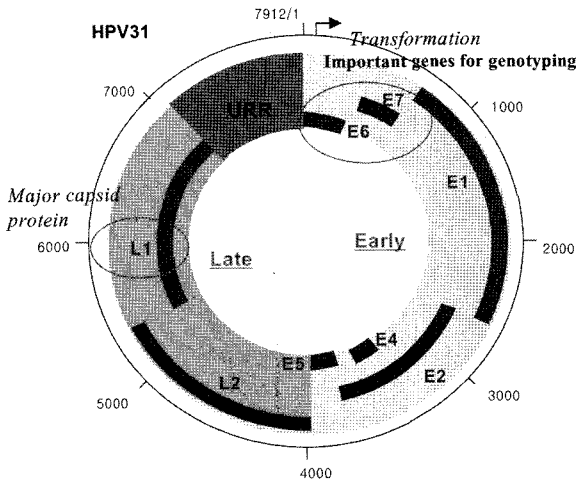

그림 11 유전자지도의 예(HPV 타입-31) [56]

문서정보가 없을 경우, **서열만으로 위험군을 판별할 수 있는 예측 방법 역시 개발되었다[56]. 이 분류시스템은 SVM에 사용되는 문자열커널(string kernel)에 기반**하고 있으며, HPV의 새로운 변종이나 다른 바이러스에 대해서도 위험도 여부를 파악할 수 있는 하나의 방법이 될 수 있다.

5. 결 론

지난 수십년간 발전해 온 생명공학 기술은 생명과학 전 분야에 걸쳐 대규모의 데이터를 생성하고 있다. 이러한 대량의 바이오데이터는 기존의 생물학 실험으로 얻을 수 있었던 자료로는 상상할 수도 없는 시야를 제공해 줄 수 있는 가능성과 함께 대량의 데이터를 정확하고(accurate) 효율적으로(efficient) 지능적(intelligent)으로 분석하는 새로운 도전과제를 제시하고 있다.

기계학습의 측면에서 볼 때에 생명과학은 새로운 하나의 실용적인 응용 분야이며, 이는 동시에 새로운 학습기법과 알고리즘 개발을 유도하는 자극제 역할도 한다. 예를 들어, 대규모의 학습 데이터를 효율적으로 학습하기 위한 새로운 커널 학습알고리즘을 연구할 필요성이 있거나 또는 다양한 이종 데이터를 통합 분석하기 위한 확률그래프모델링 기법을 개발할 필요가 있는 것이 그러한 예이다.

본고에서 살펴 본 바와 같이 다양한 **기계학습 기법들이 생명과학 연구에 활용되고 있으며, 향후 생명과학의 발전에 큰 영향을 미칠 것으로 기대된다. 생명과학의 발전은 궁극적으로는 다양한 생물의 신경계 및 지능에 대한 지식 등의 획득을 통하여 기계학습 및 인공지능 분야에도 큰 영향을 줄 것으로 예상된다.**

참고문헌

[1] Hwang, K.-B., Cho, D.-Y., Park, S.-W., Kim, S.-D., and Zhang, B.-T., Applying machine learning techniques to analysis of gene expression data: cancer diagnosis, Methods of Microarray Data Analysis(Proceedings of CAMDA 2000), Lin, S.M. and Johnson, K.F.(eds.), pp. 167-182, Kluwer Academic Publichers, 2002.

[2] Zhang, B.-T. and Hwang, K.-B., Bayesian network classifiers for gene expression analysis, A Practical Approach to Microarray Data Analysis, Berrar, D.P., Dubitzky, W., and Granzow, M.(eds.), pp. 150-165, Kluwer Academic Publishers, 2003.

[3] Hwang, K.-B. and Zhang, B.-T., "Bayesian model averaging of Bayesian network classifiers over multiple node-orders: application to sparse datasets," IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 35(6):1302-1310, 2005.

[4] Raval, A., Ghahramani, Z., and Wild, D.L., "A Bayesian network model for protein fold and remote homologue recognition," Bioinformatics, 18(6):788-801, 2002.

[5] Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S., and Grosse, I., "Identification of transcription factor binding sites with variable-order Bayesian networks," Bioinformatics, 21(11): 2657-2666, 2005.

[6] Troyanskaya O.G., Dolinski K., Owen A.B., Altman R.B., and Botstein D., "A Bayesian framework for combining heterogeneous data sources for gene function prediction(in S. cerevisiae)," Proc Natl Acad Sci,100(14): 8348-53, 2003.

[7] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D., "Support vector machine classification and validation of cancer tissue samples using microarray expression data," Bioinformatics, 16(10):906-914, 2000.

[8] Leslie, C., Eskin, E. and Noble, W. S., "The spectrum kernel: A string kernel for SVM protein classification," Proceedings of the Pacific Symposium on Biocomputing, 564-575, 2002.

[9] Vert, J.-P., "A tree kernel to analyze phylogenetic profiles," Bioinformatics, 18: S276-S284, 2002.

[10] Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T., "Protein homology detection using string alignment kernels," Bioinformatics, 20:1682-1689, 2004.

[11] Kashima, H., Tsuda, K., and Inokuchi, A., "Marginalized Kernels Between Labeled Graphs," In Proc. 20th International Conference on Machine Learning(ICML 2003), Washington, DC USA, 2003.

[12] Kim, S.-K., Nam, J.-W., Rhee, J.-K., Lee, W.-J., and Zhang, B.-T., "miTarget: microRNA target-gene prediction using a Support Vector Machine," BMC Bioinformatics, 7(1):411, 2006.

[13] Hochreiter, S. and Obermayer, K., "Kernel Methods in Computational Biology, chapter Gene Selection for Microarray Data," 319-356. Eds.: Schölkopf B., Tsuda, K. and Vert, J.-P., MIT Press, Cambridge, Massachusetts, 2004.

[14] Kim, B.H., Park, S.B., and Zhang, B.-T., "PromSearch: a hybrid approach to human core-promoter prediction," Lect. Notes Comput. SC., 3177:125-131, 2004.

[15] Tavazoie, S., Hughes, J., Campbell, M., Cho, R.J., and Church, G.M., "Systematic determination of genetic network architecture," Nat. Genet., 22:281-285, 1999.

[16] Dembele, D. and Kastner, P., "Fuzzy c-means method for clustering microarray data," Bioinformatics, 19: 973-980, 2003.

[17] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky. E., Lander, E.S., and Golub, T.R., "Interpreting patters of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," PNAS, 96:2907-2912, 1999.

[18] Mahony, S., Golden, A., Smith, T.J., and Benos, P.V., "Improved detection of DNA

motifs using a self- organized clustering of familial binding profiles," Bioinformatics, 21: i283-i291, 2005.

[19] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., "Cluster analysis and display of genome-wide expression patterns," PNAS, 95:14863-14868, 1998.

[20] Jothi, R., Zotenko, E., Tasneem, A., and Przytycka, T. M., "COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations," Bioinformatics, 22(7): 779-788, 2006.

[21] Qin, J., Lewis, D., and Noble,W., "Kernel hierarchical gene clustering from microarray gene expression data," Bioinformatics, 19: 2097-2104, 2003.

[22] Model, F., König, T., Piepenbrock, C., and Adorjan, P., "Statistical process control for large scale microarray experiments," Bioinformatics, 18:S155-S163, 2002.

[23] Ji, Y., Wu, C., Liu, P., Wang, J., and Coombes, K.R., "Applications of beta-mixture models in bioinformatics," Bioinformatics, 21(9):2118-2122, 2005.

[24] Mayrose, I., Friedman, N., and Pupko, T., "A Gamma mixture model better accounts for among site rate heterogeneity," Bioinformatics, 21:ii151-ii158, 2005.

[25] Madeira, S.C. and Oliveira, A.L., "Biclustering algorithms for biological data analysis: a survey," IEEE Transactions on Computational Biology and Bioinformatics, 1(1):24-45, 2004.

[26] Friedman, N., Linial, M., Nachman, I., and Pe'er, D., "Using Bayesian Network to Analyze Expression Data," J. Computational Biology, 7:601-620, 2000.

[27] Perrin, B.-E., Ralaivola, L., Mazurie, A., et al., "Gene networks inference using dynamic Bayesian networks," Bioinformatics, 19:ii138-ii148, 2003.

[28] Chang, J.-H., Hwang, K.-B., O, S.J., and Zhang, B.-T., "Bayesian network learning with feature abstraction for gene-drug dependency analysis," Journal of Bioinformatics and Computational Biology, 3(1):61-77, 2005.

[29] Pournara, I.V. and Wernisch, L., "Reconstruction of gene networks using Bayesian learning and manipulation experiments," Bioinformatics, 20(17): 2934-2942, 2004.

[30] Werhli, A.V., Grzegorczyk, M., and Husmeier, D., "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and Bayesian networks," Bioinformatics, 22(20):2523-2531, 2006.

[31] Segal, E., Wang, H., and Koller, D., "Discovering molecular pathways from protein interaction and gene expression data," Bioinformatics, 19 Supple 1:i264-71, 2003.

[32] Nariai, N., Kim, S., Imoto, S., et al., "Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks," In Proceedings of the 9th Pacific Symposium on Biocomputing, 336-347, 2004.

[33] Joung, J.-G., Shin, D., Seong, R.-H., and Zhang, B.-T., "Identification of Regulatory Modules by Co-clustering Latent Variable Models: Stem Cell Differentiation," Bioinformatics, 22(16): 2005-2011, 2006.

[34] Hwang, K.-B., Kim, B.-H., and Zhang, B.-T., "Learning hierarchical Bayesian networks for large-scale data analysis," Lect. Notes Comput. SC., 4232:670-679, 2006.

[35] Zhang, B.-T, Yang, J., and Chi, S.W., "Self-organizing latent lattice models for temporal gene expression profiling," Mach. Learn., 52(1/2):67-89, 2003.

[36] Whitney, A. R., Diehn, M., Popper, S. J., Alizadeh, A. A., Boldrick, J.C., Relman, D. A., and Brown, P. O., "Individuality and variation in gene expression patterns in human blood," PNAS, 100:1896-1901, 2003.

[37] Borovecki, F., Lovrecic, L., Zhou, J., Jeong, H., Then, F., Rosas, H.D., Hersch, S.M., Hogarth, P., Bouzou, B., Jensen, R.V., and Krainc, D., "Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease," PNAS, 102(31):11023-8, 2005.

[38] 김병희, 김성천, 장병탁, 기계학습에 의한 압타머칩 데이터 기반 심혈관 질환 단계의 예측, 한국컴퓨터종합학술대회 2006 논문집, 제33권 1(A), pp. 85-87, 2006.06.
[39] 엄재홍, 김병희, 이제근, 허민오, 박영진, 김민혁, 김성천, 장병탁, AptaCDSS-압타머칩을 이용한 심혈관질환 질환단계 예측 및 진단의사결정지원시스템, 한국정보과학회 가을학술발표 논문집, 제33권 2(A), pp. 28-32, 2006.
[40] 김병희, 김성천, 장병탁, Potential SVM을 이용한 압타머칩에서의 바이오마커 탐색, 한국정보과학회 가을학술발표 논문집, 제33권 2(A), pp. 22-27, 2006.
[41] Silva, J.M., et al., "Second-generation shRNA libraries covering the mouse and human genomes," Nat Genet, 37(11):1281-8, 2005.
[42] Chang, K., Elledge, S.J., and Hannon, G. J., "Lessons from Nature: microRNA-based shRNA libraries," Nat Methods, 3(9): 707-14, 2006.
[43] Saetrom, P., "Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming," Bioinformatics, 20(17):3055-63, 2004.
[44] Hayashita, Y., et al., "A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation," Cancer Res, 65(21):9628-32, 2005.
[45] Eder, M. and Scherr, M., "MicroRNA and lung cancer," N Engl J Med, 352(23):2446-8, 2005.
[46] Esquela-Kerscher, A. and Slack, F.J., "Oncomirs-microRNAs with a role in cancer," Nat Rev Cancer, 6(4):259-69, 2006.
[47] Gregory, R.I. and Shiekhattar, R., "MicroRNA biogenesis and cancer," Cancer Res, 65(9): 3509-12, 2005.
[48] Johnson, S.M., et al., "RAS is regulated by the let-7 microRNA family," Cell, 120(5): 635-47, 2005.
[49] Croce, C.M. and Calin, G.A., "miRNAs, cancer, and stem cell division," Cell, 122(1): 6-7, 2005.
[50] Huang, J.C., Morris, Q.D., and Frey, B.J., "Detecting microRNA targets by linking sequence, microRNA and gene expression data," in Tenth Annual International Conference on Research in Computational Molecular Biology(RECOMB). Venice, Italy, 2006.
[51] Joung, J.-G., Hwang, K.-B., Nam, J.-W., Kim, S.-J., and Zhang, B.-T., "Discovery of microRNA-mRNA modules via population-based probabilistic learning," Bioinformatics, 2007(in print).
[52] zur Hausen, H. "Papillomaviruses causing cancer: evasion from host-cell control in early events in carcinogenesis," Journal of National Cancer Inst., 92:690- 698, 2000.
[53] IARC Monographs on the Evaluation of the Carcinogenic Risks to Humans. Lyon, France: IARC Scientific Publications, 1995.
[54] Janicek, M.F. and Averette, H.E., "Cervical cancer: prevention, diagnosis, and therapeutics," A Cancer Journal for Clinicians, 51, 92-114, 2001.
[55] Park, S.-B., Hwang, S., and Zhang, B.-T., "Mining the risk types of human papillomavirus(HPV) by AdaCost," Lecture Notes in Computer Science, 2736: 403-412, 2003.
[56] Joung, J.-G., O, S.J., and Zhang, B.-T., "Protein sequence-based risk classification for human papillomaviruses," Computers in Biology and Medicine, 36:656-667, 2006.

황 규 백

1997.2 서울대학교 컴퓨터공학과(학사)
1999.2 서울대학교 컴퓨터공학과(석사)
2005.8 서울대학교 컴퓨터공학부(박사)
2003.12~2004.6 Harvard Medical School Children's Hospital Informatics Program(CHIP) 객원연구원
2005.8~2006.2 서울대학교 컴퓨터연구소 박사 후 연구원(바이오지능연구실)
2006.3~현재 숭실대학교 컴퓨터학부 전임강사
관심분야 : Machine Learning, Bioinformatics, Natural Language Processing
E-mail : kbhwang@ssu.ac.kr

정 제 균

2001~2006 서울대학교 바이오정보기술연구센터(CBIT) 연구원
2004.2 서울대학교 생물정보학(석사)
2007.2 서울대학교 생물정보학(박사)
2007.3 Cornell University 박사후연구원
관심분야 : Bioinformatics, Evolutionary Clustering Algorithms
E-mail : jgjoung@bi.snu.ac.kr

남 진 우

2001.2 연세대학교 생물학과(학사)
2002.8~현재 서울대학교 바이오정보기술연구센터(CBIT) 연구원
2004.8 서울대학교 생물정보학(석사)
2004.9~현재 서울대학교 생물정보학 협동과정 박사과정
관심분야 : Computational Genomics of microRNAs
E-mail : jwnam@bi.snu.ac.kr

김 병 희

2003.2 서울대학교 컴퓨터공학부(학사)
2003.3~현재 서울대학교 전기컴퓨터공학부 석박사통합과정
관심분야 : Probabilistic Graphical Models, Microarray Analysis, Bioinformatics
E-mail : bhkim@bi.snu.ac.kr

이 제 근

2004.2 고려대학교 컴퓨터학과(학사)
2004.3~현재 서울대학교 바이오정보기술연구센터(CBIT) 연구원
2004.3~현재 서울대학교 생물정보학 협동과정 석박사통합과정
관심분야 : Predictive Modeling of Gene Regulation by Proteins and/or microRNAs
E-mail : jkrhee@bi.snu.ac.kr

장 병 탁

1986.2 서울대학교 컴퓨터공학과(학사)
1988.2 서울대학교 컴퓨터공학과(석사)
1992.7 독일 Bonn 대학교 컴퓨터과학(박사)
1992.8~1995.8 독일국립정보기술연구소(GMD) 연구원
1995.9~1997.2 건국대학교 컴퓨터공학과 조교수
1997.3~현재 서울대학교 컴퓨터공학부 교수, 생물정보학, 뇌과학, 인지과학 협동과정 겸임교수
2001.1~현재 바이오정보기술연구센터(CBIT) 센터장
2002.6~현재 과학기술부 바이오지능 국가지정연구실 실장
2003.8~2004.8 MIT Computer Science and Artificial Intelligence Laboratory(CSAIL) 방문교수
2005.12~2006.2 독일 Bernstein Center Berlin 과학재단 방문교수
관심분야 : Biointelligence, Probabilistic Models of Learning and Evolution, Molecular/DNA Computation
E-mail : btzhang@bi.snu.ac.kr