바이오 데이터 마이닝을 위한 기계학습 기법

서울대학교 김성동 · 장병탁*

1. 서 론

생명체의 기본 단위인 유전자[1](gene)와 단백질[2](protein)은 생물학적 서열[3](biological sequence)로 표현될 수 있다. 많은 연구에 의해 수많은 생물학적 서열 데이터가 수집되어 데이터베이스로 구축되었다. 최근에는 축적된 데이터베이스를 관리하고 분석함으로써 유용한 지식을 얻어내어 생명 과학의 여러 문제에 활용하고 있다. 즉 축적된 데이터의 분석을 통해 DNA[4] 서열 분류, DNA 서열에서의 단백질 코딩 영역 분류, 분자의 구조와 기능의 예측, 진화 역사의 구축 등의 생명 현상 탐구가 가능하게 되었다. 생물정보학(bioinformatics)은 생명 과학과 컴퓨터 과학의 경계 부분에 위치하는 새로운 연구 분야로서 여러 종류의 다양한 생물학적 정보(biological information)에 관한 데이터베이스를 컴퓨터 상에 구축하여 이를 분석하고 응용하는 것을 목표로 한다.

생물학적 데이터베이스는 서로 다른 기관에서 수많은 사람들에 의해 구축되었기 때문에 데이터가 중복될 수 있고 이로 인한 오류를 포함할 수 있다. 또한 데이터가 증가할수록 데이터가 여러 곳에 흩어져 저장되는 단점이 있다. 따라서 데이터베이스를 통합하거나 다양하고 이질적인 데이터베이스로부터 자동적으로 추출된 선택된 정보만을 포함하는 데이터베이스의 구축이 요구된다.

이와 같은 통합된 데이터베이스의 구축을 위해 문자열 일치(string matching) 같은 전통적인 컴퓨터 알고리즘이 유용하게 사용되어 왔다. 그러나 생물학적 시스템 고유의 복잡성 그리고 데이터의 중복성과 잡음(noise)에 의한 불일치성(inconsistency)으로 인해 데이터베이스의 규모가 커질수록 기존의 알고리즘은 많은 서열 분석(sequence analysis) 문제를 해결하는데 한계를 보이고 있다.

신경망(neural network)이나 확률 그래프 모델(Markov model) 같은 인공지능의 기계학습(machine learning) 기법은 기존의 알고리즘과는 달리 많은 데이터가 주어진다면 잡음(noise)에 민감하지 않으며 중복된 정보들을 제거하고 압축하는 작업에도 적합하다. 기계학습의 기본 개념은 추론(inference) 과정이나 모델 최적화(model fitting)를 통해 데이터로부터 자동적으로 이론을 습득하거나 예제로부터 학습하는 것이다. 기계학습 기법은 대규모의 생물학적 데이터베이스를 구축하고 이의 분석을 통해 유용한 정보를 추출하는 생물정보학에서 유전자 서열과 생명체의 기능간의 근본적인 관계를 발견하는데 널리 적용되고 있다.

본 고에서는 생물정보학 분야의 여러 문제에 기계학습을 적용한 사례들을 조사하고 분석하였다. 2절에서는 생물정보학 문제를 분류하고 기계학습 기법의 역할을 살펴본다. 특히 DNA에서 RNA로 그리고 다시 단백질로 변화되는 과정에 수반되는 유전자 서열의 구조와 기능을 발견하는

*종신회원

1) 유전의 기본적인 물리적, 기능적 단위로서 세포의 활동을 조성, 관리한다
2) 하나 이상의 아미노산이 특정한 순서로 구성된 분자로서 세포 활동을 수행하며 신체 조직의 구조, 기능, 조정 등을 담당한다.
3) 단백질, DNA, RNA 분자의 구조를 선형적으로 기술한 것
4) 유전 정보를 포함하고 있는 분자로서 네 가지 기본적인 누클레오타이드의 이중 결합으로 만들어진다.

문제를 중심으로 이에 적용된 기계학습 기법을 조사, 분석하였다. 3절에서는 생물정보학 문제에 적용된 기계학습 기법을 기술하고 그 응용과 구현된 시스템을 살펴본다. 조사에 포함된 구체적인 학습 기법은 신경망 학습, 확률 그래프 모델 학습, 진화 연산, 확률 문법(stochastic grammar) 등이다. 마지막으로 4장에서는 조사된 학습 방법들의 특징들을 요약하고 향후 연구 과제에 관해 토론한다.

2. 생물정보학의 문제들과 기계학습의 역할

바이오 데이터 마이닝(biodata mininig)은 수많은 분자 생물학 연구로부터 생성되고 저장된 방대한 생물 분자 서열 데이터에서 생명체의 진화, 유전, 환경에의 적응, 학습 등의 생명 현상에 대한 지식을 얻어내는 과정 또는 그러한 지식을 얻기 위한 기술의 집합으로 정의할 수 있다. 이러한 데이터 마이닝을 통해 얻어진 지식은 신약 개발, 새로운 치료법의 개발, 예방학의 발전, 새로운 항생물질의 개발뿐만 아니라 약리학, 화학, 생태학 등의 발전에 기여할 수 있다. 본 고에서는 DNA, RNA, 단백질의 서열 데이터의 분석에 관계된 문제들을 조사하고 이들 문제에 적용된 기계학습 기법들에 대하여 살펴본다. 그림 1은 기계학습 방법을 이용한 바이오 데이터 마이닝 과정과 그 응용 분야를 간략하게 보여준다.

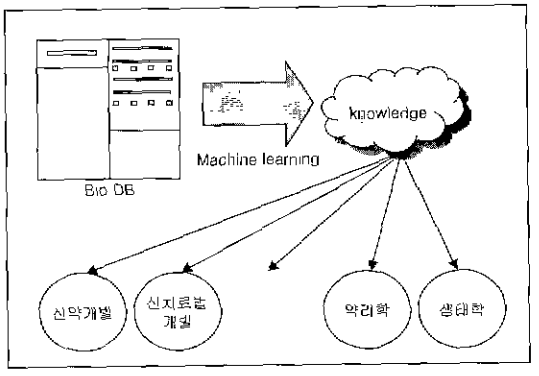

그림 1 기계학습을 이용한 바이오 데이터 마이닝과 응용

생물정보학의 주요 문제들은 그림 2와 같이 분류된다. 기계학습은 주로 서열 데이터 분석에 적용되어 왔으며, 본 고에서는 이러한 분자 서열 데이터 분석의 문제를 조사하고 이에 적용된 기계학습 기법들을 살펴본다.

분자 서열 데이터 분석이란 DNA에서 유전자를 찾고 이것이 RNA와 단백질로 변화되는 과정에 대한 연구와 그 결과물에 대한 연구를 포괄하는 용어이다. 이것은 박테리아 같은 원핵생물(prokaryotes)과 인간 같은 진핵생물(eukaryotes)에 대해서 이루어지는데 그림 3은 각각에 대한 분자 서열 데이터 분석 연구의 대상을 보여준다. 본 고에서는 생물 분자 서열 데이터 분석의 문제를 아래의 네 가지에 초점을 맞추어 조사하였으며 표 1은 이를 요약하여 보여준다.

첫째는 DNA 수준에서의 문제이다. DNA는 유전자 영역(gene region)과 비유전자 영역(intergenic region)의 연속으로 구성되는데 유전자 찾기란 주어진 서열 데이터에서 유전자 지역을 찾는 문제이다. 유전자는 생명체의 생명 활동이 일어나는 분자인 단백질로 변화되는 엑손

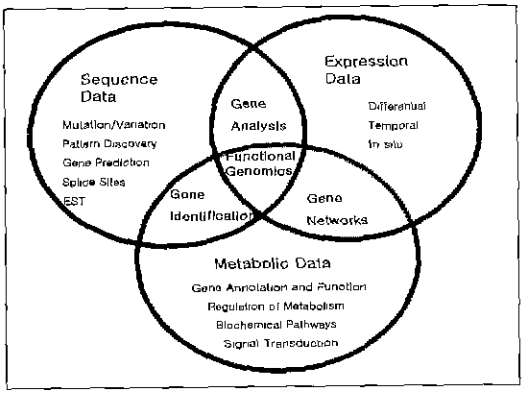

그림 2 생물정보학의 주요 문제들

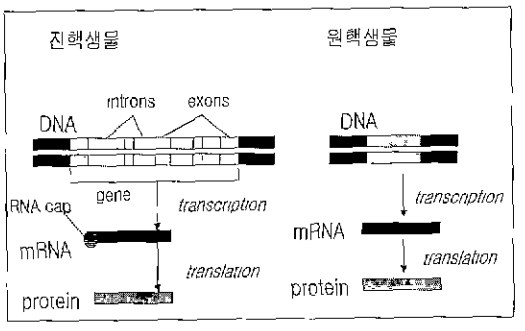

그림 3 분자 서열 데이터 분석의 대상

표 1 분자 서열 데이터 분석의 주요 문제들

	생물정보학 문제
DNA	- 유전자 찾기 • gene region • intergenic region - 유전자 구조 예측 • exon/intron/splice site • promoter/terminator • initial codon/stop codon
RNA	- 2차 구조 예측 • mRNA, tRNA, rRNA
단백질	- 단백질 구조 예측 • 1차 구조 • 2차 구조 • 3차 구조 - 단백질 기능 예측 - 단백질 군 분류
패턴 발견	- 반복 패턴의 발견 - 유사도 계산 - 의미 있는 패턴의 발견

(exon)과 그렇지 않은 인트론(intron)이 splice site를 경계로 연결되는 구조를 가진다. 또한 단백질로의 변화 과정은 전사(*transcription*)와 번역(translation)의 과정을 거치는데 유전자에서 전사의 시작 위치인 프로모터(promoter)와 끝 위치인 터미네이터(terminator), 번역 시작 위치인 시작 코돈(initial codon)과 끝 위치인 정지 코돈(*stop codon*)을 찾는 문제는 유전자 발견과 밀접한 관계가 있다. 유전자 구조 예측 문제는 유전자에서 이들 각각의 위치를 발견하는 것을 말한다.

둘째는 RNA 수준에서의 문제이다. 염기쌍(*base pair*)간의 상호작용 그리고 염기쌍과 그들의 에너지간 상호작용에서의 자유 에너지(free energy)에 기초해서 mRNA, tRNA, rRNA의 2차 구조를 계산하고 점수를 산정하는 문제를 말한다. 이러한 문제는 평가 대상이 되는 구조의 수가 매우 많기 때문에 일반적인 컴퓨터 알고리즘으로 해결하기 어려운 문제이며, 주로 신경망 학습이나 확률 문법 등이 활용된다.

셋째는 단백질 수준에서의 문제로 단백질의 구조와 기능 예측(protein structure/function prediction) 문제이다. 단백질의 기능은 3차원(3D) 구조와 관계가 있으며 유사한 구조를 가지는 단백질은 유사한 기능을 수행한다. 단백질을 구성하는 아미노산(amino acid) 서열에 의해 3D 구조가 결정되는데 아미노산 서열이 많이 다르더라도 근본적으로 다른 3D 구조의 수는 적을 수 있다. 즉 서로 다른 두 단백질의 아미노산 서열 유사도가 상대적으로 작더라도 유사한 단백질 구조와 기능을 가질 수 있다. 따라서 단백질의 기능을 예측하기 위해서는 단백질의 구조적 성질에 기초한 방법이 필요하다. 단백질의 구조는 펩티드 사슬(peptide chain)의 폴딩(folding)을 통해 생성되는데, 가능한 폴딩 패턴의 수가 무한하므로 단백질의 구조적 특성을 1차, 2차, 3차 구조의 3 단계로 구분하였다. 이렇게 함으로써 단백질의 각 구조를 이해하는 것이 보다 수월해지며 이를 통해 단백질의 기능을 예측할 수 있게 된다.

네째는 서열 데이터의 분석보다는 새로운 생물학적 패턴을 발견하는 문제이다. 이러한 패턴 발견(pattern discovery) 문제는 전체 게놈 수준에서 구조적인 조작을 찾는 문제를 포함하는데, 예를 들면 반복되는 영역의 발견, 유사도의 계산, 의미 있는 드문 패턴의 발견 등이 이러한 문제에 속한다.

생물학적 데이터베이스는 여러 기관과 수많은 사람들에 의해 구축되기 때문에 유사한 서열 데이터를 다량 포함하고 본질적으로는 데이터의 중복성과 오류를 포함하게 되는데, 기계학습 기법은 이러한 데이터베이스에 존재하는 오류를 탐지하는데 유용하다. 예를 들어, 진핵생물 유전자에 있는 잘못된 인트론 splice site를 탐지하거나 [1, 2] 포유 동물의 단백질에 있는 O-linked glycosylation 위치를 틀리게 할당한 것을 찾는데[3] 기계학습을 이용할 수 있다. 기계학습 기법은 데이터가 많이 존재한다면 데이터 오류에 민감하지 않은 특성이 있기 때문에, 오류를 포함한 불완전한 데이터가 존재하는 경우에 적용할 수 있는 장점이 있다. 또한 데이터에 의한 예측은 각각의 예로부터 주요한 특성들을 추출하고 불필요한 정보를 무시할 수 있어야 하는데, 기계학습 기법은 중복되는 분자 서열 정보를 무시하고 압축하는 작업을 효과적으로 수행할 수 있다. 예를 들어 신경망은 데이터에 공통적으로 적용할 수 있는 특성을 저장하기 위해 각각의 패턴을 위한 파라미터 대신 많은 수의 조정 가능한 파라미터를 사용한다. 이를 통해 관련된 구조와 기능을

묶을 수 있도록 복잡한 입력 분자 서열 공간의 구조를 간단한 표현으로 변형할 수 있다. 이와 같이 기계학습 기법은 분자 서열 공간에서 보다 복잡한 상호 관계를 발견할 수 있는 능력이 있다. 수많은 데이터로 표현되는 시스템을 이해하기 위해서는 정보 감축(information reduction)이 중요한데 정보 감축의 주요한 수단인 데이터의 분류나 예측을 기계학습을 통해 효과적으로 할 수 있기 때문에 기계학습 기법은 모든 자세한 것들을 포함하고 있는 원래의 데이터보다 훨씬 강력하고 유용하다.

위와 같은 장점들 때문에 생물정보학의 여러 문제에 기계학습이 적용되어 왔으며 본 고에서는 사례들을 조사하였다. 특히 실제적으로 가장 널리 응용되고 있는 신경망, 확률 그래프 모델, 진화 연산, 확률 문법에 초점을 맞추어 조사하였으며 그 결과는 표 2에 요약되어 있다.

표 2 기계학습 기법과 적용 사례들

	적 용 사 례
신경망	- 유전자 구조 예측 • exon/splice site • promoter • initial codon - 단백질의 2차 구조 예측 - RNA의 2차 구조 예측
확률 그래프 모델	- 유전자 구조 예측 • exon/splice site • promoter • initial codon - 단백질의 2차 구조 예측 - 단백질 군 분류 - 새로운 바이오 패턴 발견
진화 연산	- 단백질의 2차 구조 예측 - 단백질의 군 분류 - 새로운 바이오 패턴 발견
확률적 문법	- 유전자 구조 예측 • exon 찾기 - RNA의 2차 구조 예측

3. 기계학습을 이용한 바이오 데이터 마이닝

3.1 신경망(neural networks)

신경망은 서로 연결된 뉴런(neuron)으로 구성된 생물학적 학습 시스템을 모방한 것이다. 이것은 학습 데이터에 있는 에러에 민감하지 않기 때문에 학습 예제로부터 함수를 학습하는 일반적이고 실용적인 방법이라 할 수 있다. 필기체 문자의 인식, 음성 인식, 사람의 얼굴 인식 등의 실제적인 문제에 적용되어 왔으며 매우 효과적인 학습 방법으로 알려져 있다. 그림 4는 신경망의 일반적인 구조를 보여준다[5].

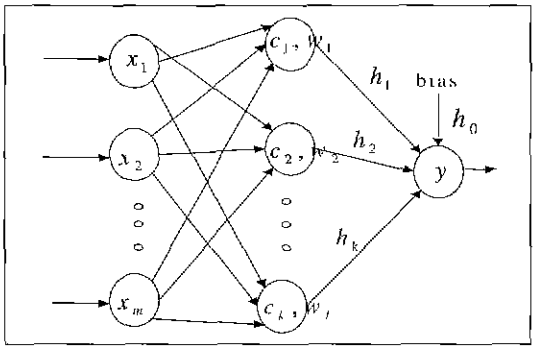

그림 4 신경망의 일반적 구조

신경망 학습을 바이오 데이터 분석에 적용한 것은 1980년대 초반부터이다. 1982년에 [4]에서는 퍼셉트론(perceptron) 학습을 이용하여 DNA 번역 시작 위치인 리보좀(ribosome) 결합 위치(초기 코돈)를 찾았다. 은닉 유닛을 가지지 않은 퍼셉트론이 일반화 능력을 가졌고 학습 데이터에 포함되지 않은 서열 데이타에서 번역 시작 위치를 찾을 수 있다는 것을 보였다. 퍼셉트론 학습은 사람에 의해 구축된 규칙보다 *E. coli* 번역 시작 위치 찾기에 더 좋은 결과를 보였다[5]. 그러나 이러한 선형 망 구조는 많은 분자 서열 인식에는 충분하지 않았다.

신경망이 서열 데이터 분석 문제에 많이 적용된 것은 다층 퍼셉트론을 위한 역전파(backpropagation) 학습 알고리즘이 사용된 이후이다. Qian과 Sejnowski는 영어 알파벳의 입력으로부터 음성 합성에 필요한 음소들을 예측하는 NetTalk 다층 퍼셉트론 구조를 적용하여 단백질의 2차 구조를 예측하였다[6]. 이 구조의 알파벳

5) $x_1, \ldots, x_m$은 입력 노드, $c_1, \ldots, c_k$는 은닉 노드, y는 출력 노드이다 그리고 $w_1, \ldots, w_k$는 각 은닉 노드의 가중치 값이다.

입력을 아미노 산이나 뉴클레오타이드(nucleotide)의 알파벳으로 바꾸고 예측되는 음소를 단백질의 2차 구조에 해당하는 helices, sheets, coil 등이나 binding sites, cleavage sites, residues 등을 나타내는 기능적인 부류로 바꿈으로써 분자 서열 데이터 분석에 적용하였다.

전사 과정의 시작 위치인 프로모터의 인식을 위해 신경망을 사용한 대표적인 예는 KBANN (*Knowledge-Based Artificial NeuralNetwork*) 시스템이다[7]. 이것은 신경망 학습과 기호 학습(symbolic learning)을 결합한 것으로 신경망의 초기 구조와 가중치 값을 결정하기 위해 명제 규칙(propositional rules)을 사용한다. KBANN의 특징은 학습시에 데이터뿐만 아니라 문제와 관련된 지식을 사용할 수 있기 때문에 신경망이 훨씬 더 빠르게 학습할 수 있고 보다 일반화 능력이 좋은 해를 찾을 수 있다는 것이다.

*Uberbacher*와 *Mural*은 신경망 학습을 적용하여 진핵생물 DNA에서의 코딩 영역인 엑손을 인식하였다[8]. 이것은 코딩 인식 모듈(CRM: *coding recognition module*)이라 불리며 GRAIL[6][9]이라 불리는 자동 서열 데이타 분석 시스템의 한 부분이다. [10]에서는 이웃하는 코돈[7](codon)의 결합 확률을 이용하여 엑손 인식 신경망의 성능 향상을 도모하였다. 즉 퍼셉트론 학습시 코돈의 종속성을 특성으로 표현하고 신경망이 이러한 특성을 고려하여 엑손을 인식할 수 있게 함으로써 코돈 자체만의 특성을 이용하여 학습하였을 때 보다 퍼셉트론의 일반화 능력을 한층 향상시켰다.

Lapedes는 신경망을 이용하여 splice site를 찾는 방법을 제안하고 결정 트리(decision tree)와 k-최근접 학습을 이용한 방법 등의 결과와 비교하였다[11]. 신경망을 이용한 방법이 다른 방법에 비해 우수한 인식의 정확도를 보였는데 91%의 acceptor[8] 인식률과 95%의 donor[9] 인식률을 보였다. NetGene 방법은 splice site와 엑손의 상호 보충성을 신경망을 통해 학습하여 splice site와 엑손의 인식을 동시에 한다[12]. 초기 버전은 사람의 유전자 서열만을 가지고 학습하였고 1992년에 인터넷을 이용해서 사용할 수 있게 되었다. 이 방법은 코딩/비코딩 영역의 예측을 하는 하나의 전체적인 망이 donor, acceptor의 임계값 할당을 위한 두 개의 지역적인 망을 조정하는 구조를 가진다.

이와같이 신경망 학습은 단백질의 2차 구조 예측 문제[13, 14], 리보좀, 프로모터, 엑손, splice site 등의 유전자 구조 발견 문제에 다양하게 적용되어 왔으며 신경망 학습의 과적합(overfitting) 문제를 고려한 구조 설계에 대한 연구[15, 16]도 진행되는 등 많은 연구가 진행되고 있다.

3.2 확률 그래프 모델(Markov model)

확률 그래프 모델은 상태(state) 집합, 기호(symbol) 알파벳 집합, 전이확률(transition probability), 발산확률(emission probability) 등으로 구성되는 확률적인 생성 모델이다. 그림 5는 시작상태(s), 종료상태(E), 삭제상태(d_i), 주상태(m_i), 삽입상태(i_i) 등으로 구성된 생물 분자 서열 분석을 위한 간단한 확률 그래프 모델을 보여준다.

확률 그래프 모델에 사용되는 기호 알파벳을 생물학적 서열에 해당하는 기호로 바꿈으로써 서열 데이터 분석 문제에 적용할 수 있다. 예를 들어 단백질을 위해 20개의 아미노 산 알파벳을 도입하거나 DNA/RNA 문제에 적용하기 위해서

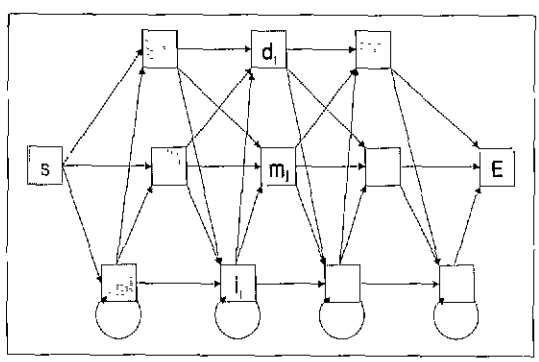

그림 5 간단한 확률 그래프 모델의 예

6) 분자 서열 분석을 통해 코딩 영역뿐만 아니라 splice junction과 번역 시작 위치도 인식한다.
7) 단백질로 변화되는 누클레오타이드의 단위로서 A, C, G, T 네 가지 중 3개의 결합이 하나의 코돈을 형성한다.
8) 인트론과 엑손의 경계
9) 엑손과 인트론의 경계

는 4개 문자의 누클레오타이드 알파벳을 이용하는 등 문제에 따라 서로 다른 기호 알파벳 집합을 정의함으로써 서열 데이터 분석 문제에 적용할 수 있다.

HMM(hidden Markov model)은 DNA에서 엑슨과 인트론을 찾거나[17] 단백질 군(protein family) 분류[18] 등에서 이용되어 왔지만 본격적으로 바이오 데이터 마이닝 분야에서 활용된 것은 1990년대 중반부터이다. 1990년대 들어 게놈 프로젝트, EST(expressed sequence tags)[19] 프로젝트들에서 얻어진 데이터들은 대부분 분자 서열의 일부분에 해당하는 데이터들이고 점점 더 이러한 데이터들이 차지하는 비중이 높아지고 있다. 이러한 데이터에 대해서는 일반적인 쌍-비교(pairwise comparison) 방법은 유사한 서열을 찾는데 한계가 있기 때문에 다중 정렬(multiple alignment)에 기반한 일치 모델(consensus model)을 이용하여 데이터베이스 탐색의 정확도나 속도를 개선하는 것이 최근의 추세이다[20, 21, 22]. 프로파일[23], 유동적 패턴[21], 블록[22] 등과 같은 서열 일치 기술은 HMM의 특별한 경우로 볼 수 있다. 그래서 1990년대 중반들어 HMM이 단백질 군이나 DNA의 모델링, 분석, 정렬 등을 위해 다른 기계학습 기법과 결합되어 본격적으로 적용되기 시작하였다.

원핵생물의 유전자를 찾기 위해 Glimmer에서는 언어와 음성 처리에서 기원한 interpolated 확률 그래프 모델을 적용하였다[24]. DNA를 구성하는 4개의 베이스(base) 각각의 확률을 기술하는 확률 그래프 모델을 구축하여 박테리아에 있는 유전자의 98%를 찾는데 성공하였다.

진핵생물이 원핵생물 유전자보다 코딩 영역의 밀도가 훨씬 높고 인트론이라는 RNA가 존재하기 때문에 진핵생물에서 유전자 찾기는 훨씬 어려운 일이다. 진핵생물, 특히 인간의 DNA에서 유전자를 찾는 것 중에서 가장 성능이 좋은 것은 HMM을 이용한 것들이다. GENSCAN은 인간의 DNA에서 유전자를 찾는데 그 중 splice site를 찾는데 매우 높은 정확도를 보인다[25]. Splice site 인식 알고리즘은 서로 다른 베이스 위치간의 연관성을 이용하는 확률적인 결정 트리인데 트리의 단말 노드에 있는 각 위치의 점수를 결정하는데 확률 그래프 모델을 이용한다. GENSCAN은 엑슨의 78%를 인식하고 81%의 엑슨 예측 정확도를 보이는데 다른 유전자 인식 시스템보다 가장 좋은 성능을 보인다. HMMgene 역시 HMM을 이용하여 인간의 유전자를 찾는 시스템이다[26]. 그리고 Genie는 generalized HMM을 이용하여 인간의 유전자 데이터를 이용하여 학습되었으며 다중 엑슨 유전자를 찾는데 적용된다[27, 28]. 이것은 69%의 유전자를 인식하고 70%의 엑슨 예측 정확도를 보인다.

또한 HMM은 분류 문제에 적용되어 단백질의 군 분류에 적용되었다. 각 군에 대한 학습 데이터를 이용하여 각 군을 위한 모델을 학습시키는 방법을 이용한다. 이를 이용하여 [29]에서는 글로빈(globin)의 하위 군을 분리하였다. 또한 하나의 단백질 군마다 하나의 HMM을 이용하는 전역적인 단백질 분류 시스템은 단백질의 분류 뿐만 아니라 유전자 찾기, 단백질의 구조/기능 예측의 문제 해결의 유용한 보조 도구로써 널리 이용되고 있다[30].

3.3 진화 연산(evolutionary computation)

자연계의 진화 과정을 컴퓨터 상에서 모의 실험함으로써 복잡한 실세계의 문제를 해결하고자 하는 것이 진화 연산(evolutionary computation)이며 이에 기반한 학습 방식은 진화 학습(evolutionary learning)이라 불린다. 이것은 문제에 대한 가능한 해들을 염색체(chromosome)로 표현한 후, 이들의 복제, 교차, 돌연변이와 같은 유전 연산자를 적용하여 변형시키면서 최적의 해를 찾는 알고리즘이다[31].

과거에 생화학자들은 단백질의 1차 및 3차 구조를 결정하는데 사용되는 기술에 있어서 많은 에러를 내포하고 있었으며 많은 시간 및 비용을 소모하였다. 따라서 그들은 단백질 폴딩(folding) 문제에 관심을 가지게 되었다. 단백질 폴딩 문제는 단백질의 1차 구조로부터 3차 구조를 예측하는 작업을 포함하는데 3차 구조는 1차 구조에 의해서 결정되며 1차 구조는 복잡하지 않아서 3차 구조에 비해서 결정하기 쉽다. 또한 전체 단백질 구조인 3차 구조는 조밀하게 얽혀있는 공모양의 2차 구조들로 구성되어 있다. 그래서 2차 구조는 3차 구조의 서브구조라고 할 수 있고 일반적으로

α-helix 또는 β-sheet로 기술된다. 여기서 유전자 프로그래밍(genetic programming)은 중간 과정으로 단백질의 1차 구조에서 2차 구조를 유도하기 위해서 사용되었다[32]. 즉 유전자 프로그래밍은 다른 단백질들이 어떻게 폴딩되는지 많은 예들을 통해서 관찰하고 일반적인 원칙을 유도해 낸다.

PRINTS는 단백질 지문(fingerprints)의 약어이다. 여기서 지문이란 단백질 군의 특성을 결정짓는데 사용되는 하나의 기준이 되는 것들의 집합이라 할 수 있다. 즉 지문은 특정한 단백질 군에 속하는 서열의 여부를 판별할 수 있는 패턴들이다. 현재 이와 관련된 다양한 데이터베이스들이 존재하고 있다. 한가지 예로서 그 중에서 PROSITE 사전(dictionary)이라는 데이터베이스가 있으며 여기에는 패턴들과 일치하는 인간의 인슐린(insulin)에 대한 짧은 서열 데이타가 포함되어 있다. PROSITE의 중요한 목적은 새로운 단백질 서열을 분석하기 위한 것이고, 더욱이 새로운 서열은 어떤 군에 속하는지 정의되어야 한다. 이러한 작업들이 비효율적인 방법으로 사람에 의해 행해지고 있다. 하지만 제안된 시스템은 진화적 알고리즘을 이용하여 단백질 서열 데이터베이스로부터 단백질 군 지문을 유도를 목적으로 데이터의 자동화된 마이닝에 사용되었다[33].

[34]에서는 바이오 패턴의 발견을 위해서 유전자 프로그래밍을 적용하였다. 누클레오타이드의 패턴 인식을 위해서는 A, G, C, T의 값으로 학습 데이터를 표현하고 단백질의 패턴 인식을 위해서는 20개의 서로 다른 아미노산을 나타내는 속성으로 학습 데이터를 기술하였다. 그리고 PROSITE 패턴 언어를 이용하여 패턴을 표현하였다. 발견된 패턴의 중요도를 학습 데이터로부터 직접 계산하여 확률로 표현하고 이를 적합도 함수에 반영함으로써 임의 패턴인지 기능적으로 관련된 패턴인지를 판단할 수 있다. 이러한 패턴 인식 방법은 단백질 군간의 구분을 위해서도 사용될 수 있다.

3.4 확률 문법 (stochastic grammar)

생물 분자 서열에 내재한 정보의 분석은 언어학(linguistics)에서의 언어 분석과 많은 유사성을 가진다. 형식 문법(formal grammar)은 원래 자연언어를 기술하기 위해 개발되었으며 컴퓨터 언어나 컴파일러의 분석과 설계에 널리 이용되어 왔다. 분자 생물학의 많은 문제들이 형식 언어로 표현될 수 있으므로[35] 최근에는 생물학적인 서열 분석에 형식 문법을 적용하여 유전자 구조 예측이나 RNA 구조 예측을 한다. 그림 6은 간단한 문맥 자유 문법과 그로부터 유도된 특정한 서열의 한 예를 보여준다[36].

GENLANG은 유전자 구조 예측의 문제에 언어학적인 방법을 적용한 것이다[37]. 여기서는 진핵생물에서 엑손과 tRNA를 찾기 위한 문법을 설계하고 파서를 개발하였다. 유전자의 구조와 생물 분자 서열의 특성들을 형식 문법을 이용하여 기술하고 구조적인 패턴 인식(syntactic pattern recognition) 방법을 이용하여 구조들을 찾는다. 유전자 문법은 프롤로그(prolog) 언어를 이용하여 구현된 유한 구절 문법(DCG, definite clause grammar)이고 속도의 개선을 위해 차트 파싱 방법을 이용한다.

학습된 확률적인 문법은 모듈식으로 다른 문법과 결합되어 이용되기도 하는데 tRNA 유전자 탐색을 위한 tRNA SCFG(stochastic context free grammar)는 인트론 문법과 결합된다[38].

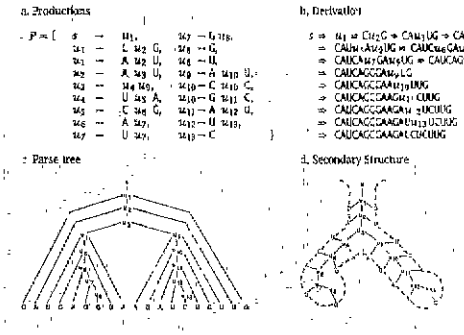

그림 6 문맥 자유 문법과 서열 유도의 예

4. 결 론

분자 생물학의 많은 연구를 통해 구축된 생물학적 데이터를 효과적으로 조직, 저장, 관리하고 그것으로부터 필요한 지식을 얻어내는 생물정보학 분야에서 유전자 구조 예측, 단백질 구조/기능의 예측, 단백질의 분류, 새로운 생물학적 패턴

의 발견 등의 문제에 다양한 기계학습 기법이 적용되어 왔다. 본 고에서는 기계학습 기법 중 신경망, 확률 그래프 모델, 진화 학습, 확률 문법 등의 방법이 생물정보학 문제 해결에 적용된 사례들을 조사하였다.

본 조사를 통해서 기계학습 기법이 분자 생물 데이터 분석 및 예측의 다양한 분야에 걸쳐 폭넓게 활용되고 있음을 알 수 있었다. 이와 같이 기계학습은 데이터베이스로부터 필요한 지식을 추출하는데 뿐만 아니라 커다란 데이터베이스가 필연적으로 수반하게 되는 오류 등을 검출하는 데에도 활용되고 또한 유사한 데이터들을 군집화하고 집단간의 거리 계산을 함으로써 용량이 큰 데이터베이스를 다룰 수 있게 하는 수단도 제공한다.

21세기는 인간의 건강, 유전, 생명 등에 대한 이해에 획기적인 변화를 가져올 것이다. 이러한 변화는 모든 생명체의 게놈에 포함된 기본적인 유전 현상 등을 해독함으로써 가능해진다. 이를 위해 대규모의 데이터의 축적과 축적된 데이터로부터 생명체의 각 기능의 상호 작용에 대한 지식을 얻어내는 것이 필수적인 과제가 될 것이다. 지금까지 기계학습 기법은 위에서 살펴본 것처럼 생물정보학의 여러 분야에서 많은 결과를 보였지만 앞으로 해결해야 할 더 많은 문제들과 발견들이 남아있기 때문에 보다 지속적이고 깊은 연구가 필요할 것이다.

참고문헌

[1] S. Brunak, J. Engelbrecht, and S. Knudsen. Cleaning up gene databases. *Nature*, 343:123, 1990.

[2] S. Brunak, J. Engelbrecht, and S. Knudsen, Neural entwork detects errors in the assignment of pre-mRNA splice site. *Nucl. Acids Res.*, 18:4797-4801, 1990.

[3] J. E. Hansen, O Lund, J. Engelbrecht, H. Bohr, J. O. Nielsen, J. E-S. Hansen, and S. Brunak. Prediction of O-glycosylation of mammalian proteins: Specificity patterns of UDP-GalNAc: polypeptide n-acetylgalactosaminyltransferase. *Journal of Biochemistry and Biology*, 307:801-813, 1995.

[4] G. D. Stormo, T. D. Schneider, L. Gold and A. Ehrenfeucht, Use of the preceptron algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research*, 10(9): 2997-3011, 1982.

[5] G. D. Stormo, T. D. Schneider, L. M. Gold. Characterization of translational initiation sites in e. coli. *Nucleic Acids Research*, 10(9): 2991-2996, 1982.

[6] N. Qian and T J. Sejnowski. Predicting the secondary structure of globular proeins using neural network models. *Journal of Molecular Biology*, 202:865-884, 1988.

[7] G. Towell, J Shavlik, and M. Noordewier. Refinement of approximate domain theories by knowledge-based neural netoworks. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 861-866. AAAI Press, Menlo Park, CA, 1990.

[8] E.C. Uberbacher and R. J. Mural. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proceedings of National Academic Science, USA*, 88:11261-11265, 1991.

[9] E.C. Uberbacher, J.R. Einstein, X. Guan, and R. J. Mural. Gene recognition and assembly in the GRAIL system: Progress and challenges In H. Lim, J. Fickett, C. cantor, and R. Robbins, editors, *Proceedings of the Second International Conference on Bioinformatics. Supercomputng, and Complex Genome Analysis*, pages 465-476. World Scientific, Singapore, 1993.

[10] R. Farber, A. Lapedes, and K. Sirotkin. Determination of eukaryotic protein coding regions using neural networks

and information theory. *Journal of Molecular Biology*, 226:471-479, 1992.

[11] A. Lapedes, C. Barnes. C. Burks, R. Farber, and K. Sirotkin. Application of neural networks and other machine learning algorithms to DNA sequence analysis. In G. Bell and T. marr, editors, *Computers and DNA, SFI Studies in the Sciences of Complexity, vol. VII*, pages 157-182. Addison-Wesley. 1989.

[12] S. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology*, 220:49-65, 1991.

[13] H. Hohr, J. Bohr, S. Brunak, R. M. J. Cotterill, B. Lautrup, L. Norskov, O. H. Olsen, and S. B. Petersen. Protein secondary structures and homology by nueral networks: The α-helices in rhodopsin. *FEBS Letters*, 241:223-228, 1988.

[14] L. H. Holley and M. Karplus. Protein secondary structure prediction with a neural network. *Proceedings of Nat. Acad. Sci. USA*, 86:152-156, 1989.

[15] A. Krogh and S. K. Riis. Prediction of beta sheets in proteins. In M. C. Mozer S. Touretzky and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pp. 917-923. MIT Press, Boston, MA, 1996.

[16] S. K. Riis and A. Krogh Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments, *Journal of Computer and Biology*, 3:163-183. 1996.

[17] G. A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bull. Mathem. Biol.*, 51:79-94, 1989.

[18] J. V. White, C. M. Stultz, and T. F. Smith. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. Mathem. Biosci., 119:35-75, 1994.

[19] D. Gerhold and C. T. Caskey. It's the genes! EST access to human genome content. *Bioessays*, 18:973-981, 1996.

[20] A. Bairoch. The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucl. Acids Res.*, 21:3097-3103, 1993.

[21] G. J. Barton. Protein multiple sequence alignment and flexible pattern matching. *Meth. Enzymol.*, 183:403-427, 1990.

[22] S. Henikoff and J. G. Henikoff. Protein family classification based on searching a database of blocks. *Genomics*, 19:97-107, 1994.

[23] M. Gribskov, A. D. McLachlan, and D. Eisenberg. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, 84:4335-4358, 1987.

[24] S. Salzberg et al., Microbial Gene Identification Using Interpolated Markov Models, *Nucleic Acids Research*, Vol. 26, No. 2, 1998, pp. 544-548.

[25] C. Burge and S. Karlin, Prediction of Complete Gene Structures in Human Genomic DNA. *Journal of Molecular Biology*, Vol. 268, 1997, pp. 78-94.

[26] A. Krogh. Two methods for improving performance of an HMM and their application for gene finding In *Proc. of Fifth Int. Conf. on Intelligent Systems for Molecular Biology*, ed. Gaasterland, T. et al., Menlo Park, CA: AAAI Press, 1997, pp. 179-186.

[27] D. Kulp, D. Haussler, M.G. Reese, and F.H. Eeckman, A generalized Hidden Markov Model for the recognition of human genes in DNA, *ISMB-96*, St. Louis, MO, AAAI/MIT Press.

[28] M.G. Reese, F.H. Eeckman, D. Kulp, D. Haussler, Improved splice site detection

in Genie. *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB)* 1997, Santa Fe, NM, ACM Press, New York.

[29] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Hausseler. Hidden Markov models in computational biology: Applications to protein modeling. Journal of Molecular Biology, 235:1501-1531, 1994.

[30] E. L. L. Sonnhammer, S. R. Eddy, and R. Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins: Structure, Function, and Genetics, 28:405-420, 1997.

[31] 장병탁. 인공 진화에 의한 학습 및 최적화. *제어자동화시스템학회지*. 제1권 제3호. pp. 52-61, 1995.

[32] Simon Handley. Automated Learning of a Detector for α-Helices in Protein Sequences Via Genetic Programming. In *Proceedings of the Fifth International Conference on Genetic Algorithms*. pp. 271-278. 1993.

[33] Bjorn Olsson. Using Evolutionary Algorithms in the Design of Protein Fingerprints. In *Proceedings of the Genetic and Evolutionary Computation Conference*. pp. 1636-1642. 1999.

[34] Yuh-Jyh Hu. Biopattern Discovery by Genetic Programming. In *Proceedings of the Third Annual Genetic programming Conference*. pp. 152-157, July 22-25, 1998.

[35] V. Brendel and H. G. Busse. Genome structure described by formal languages. *Nucl. Acids Res.*, 12:2561-2568. 1984.

[36] Pierre Baldi and Soren Brunak. Bioinformatics: The Machine Learning Approach. The MIT Press, 1998.

[37] S. Dong and D. Searls, Gene Structure Prediction by Linguistic Methods, *Genomics*. Vol. 23, Oct. 1994, pp. 540-551.

[38] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucl. Acids Res.*, 22:2079-2088, 1994.

김 성 동

1991 서울대학교 컴퓨터공학과 학사
1993 서울대학교 컴퓨터공학과 석사
1999 서울대학교 컴퓨터공학과 박사
1999~현재 서울대학교 컴퓨터신기술 공동연구소 연구원
관심분야: 자연언어처리, 기계번역, 기계학습
E-mail: sdkim@scai.snu.ac.kr

장 병 탁

1986 서울대학교 컴퓨터공학과 학사
1988 서울대학교 컴퓨터공학과 석사
1992 독일 Bonn 대학교 컴퓨터과학과 박사
1988~1992 Bonn 대학교 AI Lab 연구원
1992~1995 독일국립정보기술연구소(GMD) 연구원
1995~1997 건국대학교 컴퓨터공학과 조교수
1997~현재 서울대학교 컴퓨터공학부 조교수
관심분야: 인공지능, 기계학습, 신경망, 진화연산
E-mail: btzhang@scai.snu.ac.kr