지식기반 유전자 알고리즘에서 추출된 규칙을 이용한 Cleavage Site 예측

조연진[0], 김현철
고려대학교 컴퓨터교육학과
{jx[0], hkim}@comedu.korea.ac.kr

Cleavage Site Prediction Using the Rule Extracted from Knowledge-Based Genetic Algorithm

Yeun-Jin Cho[0], Hyeoncheol Kim
Dept. of Computer Science Education, Korea University

요 약

Cleavage Site 분석 및 예측은 바이러스 증식에 필요한 핵심 단백질인 Protease($3CL^{pro}$)를 예측하게 하고, 예측한 Protease의 활성을 억제함으로써 바이러스 증식을 저지하게 된다. 본 연구에서는 신경망과 결정트리, 유전자 알고리즘을 이용하여 SARS-CoV의 cleavage site를 분석하고, 학습 결과에서 추출된 규칙(Rule)에 의해 cleavage site를 예측한다. 또한 신경망에서 학습된 지식(Knowledge)을 이용하여 유전자 알고리즘의 성능을 향상시키는 지식기반 유전자 알고리즘 (KBGA: Knowledge-Based Genetic Algorithm)을 제안한다.

1. 서 론

지금까지 인류가 접해보지 못한 변종 Coronavirus는, 광둥성 일대 야생동물에 살던 바이러스가 돌연변이를 일으켜 사람에게 옮겨 오면서 SARS 파동으로 시작되었다.

콜럼비아 게놈과학센터의 마르코 마라 박사 외 연구팀은 기존 Coronavirus와 유전적 구조에 현저한 차이를 갖는 변종 Coronavirus를 SARS의 원인균으로 지목하고 SARS 바이러스를 SARS Coronavirus (혹은 SARS-CoV)라는 제4군 Coronavirus (CoV)로 그림 1 과 같이 분류하였다[1].

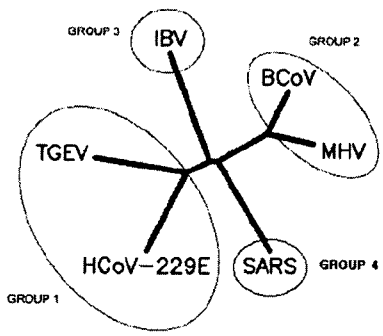

그림 1. SARS Coronavirus와 다른 Coronavirus

SARS-CoV를 무력화 시킬 수 있는 방법은 SARS-CoV의 핵심 단백질 가운데 하나인 protease ($3CL^{pro}$)의 활성을 억제하여 바이러스 증식을 방해하는 것이다. CoV를 증식 분해시키는 protease는 다른 군의 CoV protease와 서로 비슷한데, 즉 protease에 의해 분해되는 위치를 나타내는 CoV 1군의 cleavage site는 다른 2~3군의 cleavage site와 유사하다는 것이다. 따라서 1~3군의 CoV cleavage site 분석 결과는 4군의 SARS-CoV cleavage site를 예측하게하고, 향후 CoV로 변종된 바이러스 질병에 대처하게 할 것이다[2].

본 연구에서는 기계학습 알고리즘에서 bio-data의 분류기(Classifier)로 자주 사용되는 결정트리(C5.0), 신경망, 유전자 알고리즘을 이용하여 규칙(Rule)을 생성하고, 생성된 규칙에 의하여 SARS-CoV cleavage site를 예측한다. 그리고 Feed-forward neural network에서 학습된 지식(Knowledge)을 이용하여 전통적인 유전자 알고리즘에서 나타나는 문제점을 해결하는 지식기반 유전자 알고리즘(Knowledge-Base Genetic Algorithm)을 제안한다. 지식기반 유전자 알고리즘(KBGA)을 사용하여 생성된 분류 규칙은 다른 알고리즘에서 발견되지 않은 높은 성능의 새로운 규칙을 추출하였다.

2. 실험 데이터와 cleavage site의 규칙추출 방법

2.1 실험 데이터

실험에 사용한 cleavage site는 특정 protease에 의해 바이러스의 단백질이 분해되는 위치(Position)를 말하며, GenBank database에서 12개의 SARS-CoV와 다른 군의 Coronavirus 12개, 총 24개가 사용되었다[3].

각 CoV genome 서열마다 11개의 cleavage site가 포함되어 총 264(=24×11)개, 그 중 중복된 서열을 제외하고 70개의 cleavage site (Positive data)를 사용하였다.

실험 데이터 집합을 구성하는 cleavage 서열은 P4, P3, P2, P1, P1', P2', P3', P4' (8자리)로 나타내고 20개의 아미노산으로 표현한다. 모든 P1의 위치에는 아미

노산 'Q'를 포함하고 있기 때문에 Non-cleavage site는 Kiemer 외(Kiemer et al., 2004)에서 사용한 방법과 같이 coronavirus genome 서열에서 P1의 위치에 'Q'가 들어간 non-cleavage site (Negative data)를 선택하여 중복된 서열을 제외하고 1267개를 사용하였다[4].

신경망의 학습(training)을 위해 입력되는 8개의 아미노산은 각 20bit 이진수로 변환(encoding)한다. 예를 들어 Alanine(A)는 "10000000000000000000"로 변환하여, 총 160(=8×20)개의 입력(input) 노드를 갖는다[5]. 각 1, 2, 4, 6, 8의 은닉(hidden) 노드별 실험 중에서, 가장 좋은 결과를 보인 2개의 은닉 노드와 cleavage site 클래스(cleavage site: 1, non-cleavage site: 0)를 구분하는 1개의 출력(output) 노드를 갖는다.

2.2 Cleavage site의 규칙추출 방법

■ Sequence logo는 CoV genome의 엔트로피를 기반으로, 열의 전체 높이는 서열의 위치가 가지고 있는 정보의 총량을 표시하고 글자의 크기는 상대적인 발생 빈도를 나타낸다[그림 2]. 단백질 모티프와 같은 비교적 짧은 서열에 유용하고, 주어진 서열이 데이터 전체를 반영하지 않을 경우 정확한 position을 예측하기 힘들다.

Suquence logo를 통해 확인된 3개의 Consensus Pattern은 'LQ', 'LQ[S/A]', '[T/S/A]X[L/F]Q[S/A/G]'이다[4].

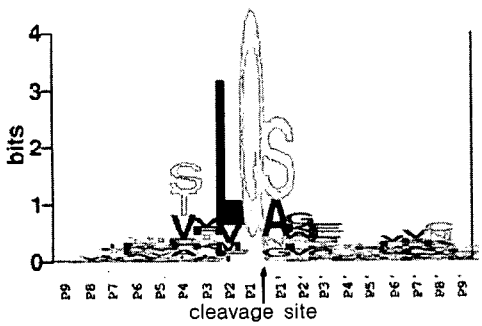

그림 2. SARS-CoV cleavage site의 Sequence logo

■ 결정트리(DT)는 획득 정보량 (information gain)이 최대값을 가지는 입력 필드를 분류 필드로 하여 트리 모형의 규칙을 만들게 되는 것으로 분류된 결과를 쉽게 이해할 수 있다. 실험에서 사용한 C5.0 알고리즘에서는 "If L@P2 ^ [A/C/G/N/S]@P1, then cleavage site"인 규칙이 생성되었다.

■ 신경망(NN)은 학습 데이터 집합을 반복적으로 처리하면서 연속적인 입력 벡터에 대한 예측값과 cleavage site 클래스 값을 비교하면서 값의 차이에 대한 평균제곱오차(MSE)가 최소가 되도록 수정한다.

Kiemer, 외[4]는 SARS-CoV cleavage site 분석을 위해 NN을 사용하였다. 그들은 Consensus pattern을 능가하는 성능을 나타내었으나, 뚜렷한 지식에서 site를 분석하는 것이 아니라 단지 cleavage site인지의 여부를 예측 하였다.

그러나 우리는 "*왜 cleavage site*"으로 예측했는지가 궁금한 것이고, 학습된 신경망에서 추출한 "*If-Then*" 규칙을 통해 알고자 하는 것이다. 신경망에서 규칙을 추출하기 위한 많은 선행연구가 진행되었으며[6], 본 실험에서는 OAS (Ordered-Attribute Search) 알고리즘을 이용하여 규칙을 추출한다[7].

■ 유전자 알고리즘(GA)은 개체 집단(population)을 분류(classifier) 규칙의 형식(아미노산과 '*': don't cares)으로 표현하고 SARS-CoV Positive 데이터와 일치하는 유전자(gene)가 염색체(chromosome) string position에 존재할 경우 염색체의 적합도(fitness)를 높여줌으로서 적합도가 높은 개체를 선택해 나간다[8]. 그러나 GA의 모든 component들이 랜덤을 기반으로 진행되기 때문에 전혀 표현될 수 없는 구조까지 나타냄으로서, 생성된 규칙의 결과는 consensus pattern 보다 더 낮은 성능을 보였다. 따라서 GA는 bio-data의 제한된 여러 요소들을 극복하기 위해, GA 실행과정에 전문이론과 경험적인 지식을 적용한 새로운 방법의 시도가 필요하다[9].

■ 지식기반 유전자 알고리즘(KBGA)은 GA의 성능을 향상시키기 위해 실행 과정에서 적용할 수 있는 이론과 지식을 이용한 기법으로 제시된다.

본 연구에서는 학습된 NN에서 추출한 지식(*If I45 and I150 then cleavage site*)을 이용하여 GA의 초기(initial) 개체 집단을 그림 3과 같이 설정한다. 이것은 유전자의 우수 형질을 확보함으로서, 최적의 값을 효율적으로 탐색하게 한다. NN에서 추출된 지식은 각 아미노산의 종류와 위치 정보를 포함하고 있어, 전통적인 GA 보다 더 빠른 실행 속도로 수렴하는 결과를 얻게 하였다.

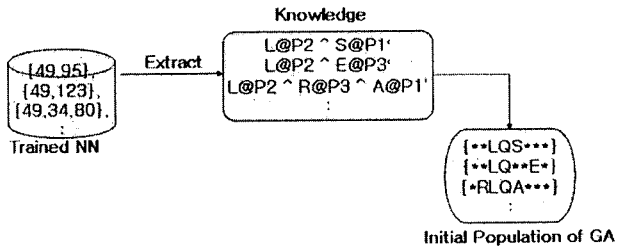

그림 3. KBGA의 흐름도

적합도 함수는 규칙의 일반화(generalization)와 정확도(accuracy)를 동시에 모두 만족시키는 조건으로 최적의 해를 검색한다.

$$f(n) = \frac{num\ of\ positive}{num\ of\ positive + num\ of\ negative + 1} \times 100 + d$$

d 는 '*'(don't care symbol)의 개수로 규칙의 일반화를 만족하게 한다. 다음 세대의 재생산을 위해서는 룰렛휠 선택과 one-point 교차(crossover), 0.01의 돌연변이(mutation) 연산을 적용하였다.

3. 실험결과 및 추출된 규칙의 성능

각 알고리즘에서 추출된 규칙의 성능은 coverage와 accuracy로 다음과 같이 계산한다.

$$Coverage = \frac{TP+FP}{TP+TN+FP+FN}$$

$$Accuracy = \frac{TP}{TP+FP}$$

TP는 실제 cleavage site와 매칭되는 규칙의 개수이고, FP는 non-cleavage site와 매칭되는 규칙의 개수이다.

모든(positive + negative) cleavage site에는 P1의 위치에 'Q'를 포함하고 있으므로, 추출된 규칙에서 'Q'를 제외한 규칙으로 성능을 평가한다.

Consensus 규칙과 DT 규칙은 높은 coverage와 낮은 accuracy를 나타내는데, 표 1에서와 같이 이것은 NN이나 KBGA에서 생성된 규칙들의 범위 안에 포함된다.

NN은 Kiemer, 외[4]의 실험과 동일하게 3-fold cross validation방법을 사용하였다. 입력노드 160개, 은닉노드 2개, 출력노드 1개의 구조로, 97.9%의 일반화 성능과 99.6%의 학습 정확률을 나타내었다.

GA에서 생성된 규칙은 기준(coverage 17% 이상, accuracy 60% 이상)에 미달되어 표 1에 포함되지 못하였다. KBGA와 비교해 볼 때 성능에 매우 큰 차이를 볼 수 있는데, 이것은 초기 집단 생성에서 우수형질을 확보했는지의 여부가 결과에 큰 영향을 준다는 것으로 확인할 수 있다.

KBGA에, NN에서 학습된 지식의 사용 외에 참조할 수 있는 그 밖의 consensus 규칙과 DT 규칙을 사용하여 실험한 결과, NN의 지식을 이용하여 생성된 규칙의 성능이 가장 우수하였다.

각 알고리즘에서 생성된 규칙의 성능과 포함 관계는 "DT<NN<KBGA"의 순서로 형성된다. KBGA에서만 발견된 *S@P4 ^ S@P1'*는 가장 높은 86.6%의 정확률 나타내고 있다.

표 1. 각 분류 알고리즘에서 추출된 규칙의 성능(coverage 17% 이상, accuracy 60% 이상)

	Positive Rules	Coverage(%)	Accuracy(%)
Consensus rules, DT(C5.0) rule	**LQS***	36.35	75.76
	LQA*	26.11	78.26
NN rules	**LQS***	36.35	75.76
	LQA*	26.11	78.26
	LQE*	21.90	71.43
	V*LQ****	20.71	60.87
	T*LQ****	20.32	77.78
	*RLQ****	17.30	85.71
KBGA rules	**LQS***	36.35	75.76
	LQA*	26.11	78.26
	LQE*	21.90	71.43
	V*LQ****	20.71	60.87
	T*LQ****	20.32	77.78
	S**QS***	18.73	86.67
	*RLQ****	17.30	85.71

4. 결 론

본 연구에서는 SARS-CoV cleavage site 예측을 위해 기계 학습의 분류 알고리즘 중, entropy 기반의 DT, attribute 기반의 NN와 optimization 기반의 GA을 이용하여 각각의 알고리즘에서 *"If condition Then cleavage site"*의 규칙을 추출하고 그 성능을 비교하였다.

NN의 규칙은 consensus pattern보다 우수한 성능을 보였으며, NN의 지식을 이용한 KBGA의 사용은 기존의 GA 성능을 보다 향상시켰다. 이것은 bio-data의 매우 독특한 특성과 여러 제한 사항들을 극복하는 알고리즘의 필요성을 나타낸다.

향후 연구 과제로는 GA 초기 집단 생성에 사용한 지식을 선택과 교차, 돌연변이 등의 다양한 연산에 적용해 봄으로서 bio-data에 더욱 특정(specific)한 지식기반 알고리즘의 연구가 필요하다.

참고문헌

[1] Marra MA, et al.: The Genome Sequence of the SARS-Associated Coronavirus. SCIENCE VOL 300, 1399-1404. (2003)

[2] Anand,K., Ziebuhr,J., Wadhwani,P., Mesters,J.R. & Hilgenfeld,R.: Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. Science 300, 1763-1767. (2003).

[3] Gaoa F, Oua HY, Chena LL, Zhenga WX, Zhanga CT.: Prediction of proteinase cleavage sites in polyproteins of coronaviruses and its applications in analyzing SARS-CoV genomes. FEBS Letters 553, 451-456. (2003)

[4] Kiemer L, Lund O, Brunak S, Blom N.: Coronavirus 3CL-pro proteinase cleavage sites: Possible relevance to SARS virus pathology. BMC Bioinformatics (2004)

[5] Narayanan, A., Keedwell, E.C. and Olsson, B.: Artificial Intelligence Techniques for Bioinformatics. Bioinformatics (2002)

[6] Fu, LiMin.: Introduction to knowledge-based neural networks. Knowledge-Based Systems 8(6), 299-300 (1995)

[7] Kim, Hyeoncheol.: Computationally Efficient Heuristics for If-Then Rule Extraction from Feed-Forward Neural Networks. Lecture Notes in Artificial Intelligence, Vol. 1967, 170-182. (2000)

[8] De Jong, K.A. and Spears, W.M.: Learning Concept Classification Rules Using Genetic Algorithms. Proceedings of the I Zth. international Conference on Artificial Intelligence. 651-656. (1991)

[9] Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press (1996)