부정 선택을 이용한 DNA의 패턴 분류

이동욱° 심귀보
중앙대학교 정보통신연구소°
중앙대학교 전자전기공학부
dwlee@wm.cau.ac.kr°, kbsim@cau.ac.kr

Classification of DNA Pattern Using Negative Selection

Dong-Wook Lee° Kwee-Bo Sim
Information and Telecommunication Research Institute, Chung-Ang University°
School of Electrical and Electronics Engineering, Chung-Ang University

요 약

인간 및 다른 생물들의 DNA 서열이 밝혀짐에 따라 DNA 서열 정보를 이용할 수 있는 계산적 처리방식에 대한 요구가 늘어나고 있다. 본 논문에서는 DNA의 패턴을 분류할 수 있는 면역계 부정 선택에 기반한 알고리즘을 제안한다. 부정 선택은 면역세포 생성시 자신을 인식하지 않는 항원 인식부를 생성하기 위한 과정이다. 이 항원 인식부를 통해 자기와 비자기를 구별한다. 이것을 n개의 자기 또는 비자기 집단으로 확장하고 n개의 항원 집단을 구성하면 n개의 패턴 분류가 가능하다. 본 논문에서는 부정 선택에 기반한 DNA 염기 레벨에서의 패턴 분류방법과 아미노산 레벨에서의 패턴분류 방법을 제안한다.

1. 서 론

분자생물학의 발달과 게놈프로젝트를 통하여 인간 및 다른 생물들의 유전자 정보를 갖는 DNA의 서열이 점차 밝혀지고 있다. 하지만 게놈의 DNA 서열을 모두 알았다고하여 이중 어떤 부분이 유전자 이고, 어떤 유전자가 언제 어떻게 발현되는지는 알 수 있는 것은 아니다. 30억 염기의 인간 유전자중 대략 10%정도만이 단백질을 합성하는 유전정보를 가지고 있는데, 이 영역도 매우 복잡하게 분포 되어있다. 따라서 본격적인 게놈 프로젝트는 밝혀진 게놈정보를 이용해 유전자의 기능을 밝히는 포스트 게놈시대로 들어섰다. 또한 포스트 게놈시대에 들어서면서 유전정보를 정보학으로 다루는 바이오정보기술(bioinformatics)의 중요성이 점점 높아지고 있다.

본 논문에서는 면역계의 면역세포 생성메커니즘 중 하나인 부정 선택(negative selection)을 이용한 DNA 패턴 분류 알고리즘을 제안한다. 생체의 면역계는 외부의 항원을 인식하기 위한 다양한 항체를 생성한다. 항체를 생성하는 대표적인 면역세포는 B-세포로서, B-세포는 자기 자신을 항원으로 인식하지 않기 위해 초기 생성시 부정 선택의 과정을 거친다. 이와 같이 부정 선택을 거친 B-세포들은 자기 자신과 외부물질을 분류하는 능력을 가진다.

생물정보학에서 패턴 분류가 이용되는 부분은 유전자 영역(gene region)과 비유전자 영역(intergenic region)의 구분, RNA의 구조 예측, 단백질 구조 예측, 단백질 군 분류, DNA 칩의 분석, 유전자 발현정보 분류 등이 있다. 이와 같은 문제를 해결하기 위해 신경회로망, 진화연산, 확률 그래프 모델 등의 기계학습법이 이용된다[1, 2]. 본 논문에서는 부정 선택에 기반한 DNA 염기 레벨에서의 패턴 분류 방법과 아미노산 레벨에서의 패턴 분류 알고리즘을 제안하고 그 유효성을 검토한다.

2. DNA의 구조

모든 생물체는 각자 고유의 DNA(deoxy-ribonucleic acid)를 가지고 있다. DNA는 개체의 특성을 발현시키는 유전코드로서, A(아데닌) T(티민, RNA에서는 U:우라실) G(구아닌) C(시토신)의 4개의 염기배열로 이루어져 있다. 그 구조는 길다란 나선형의 형태로 꼬여 있으며 A는 T, G는 C와 상보적으로 결합하고 있다. 또한 DNA는 염기 3개의 배열이 한 의미단위를 이루어 해석된다. 이 의미단위를 생물학적인 용어로 코돈(codon)이라 한다. 코돈의 가지 수는 4×4×4 = 64개이며 이것이 코드화하는 아미노산은 20가지이다. 코돈의 64가지 패턴에 대하여 생성하는 아미노산이 20가지인 이유는 다른 코돈이 같은 아미노산을 만들기도 하기 때문이다. 이것은 표 1에 나타나 있다[3].

DNA는 RNA로 전사되어 리보솜에서 단백질로 번역된다. 즉 아미노산을 암호화하는 DNA의 배열에 따라 아미노산의 합성순서를 결정하여 여러 종류의 단백질을 만들어낸다. RNA의 단백질로의 번역은 AUG에서 시작해서 UGA(UAA,UAG)에서 번역이 끝난다. 따라서 DNA 코드 중 단백질로 번역되는 부분은 시작코돈인 AUG와 종료코돈인 UGA(UAA, UAG) 사이에 존재하는 염기들이다.

그림 1은 DNA의 번역 예를 나타낸다. 하나의 코돈이 하나의 아미노산을 생성하며 시작코돈의 위치에 따라 코돈 즉 3개의 염기가 짝지어지는 방법이 바뀐다. 따라서 같은 DNA도 코돈의 시작부위에 따라 세 가지 방식으로 해석될 수 있다. 이 세 가지로 묶는 방식을 reading frame이라 한다.

표 1. RNA(DNA) 코돈과 생성하는 아미노산

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	정지	UGA	정지	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU		CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

아미노산 약어 알라닌-Ala, 아르기닌-Arg, 아스파라긴-Asn, 아스파르트산-Asp, 시스테인-Cys, 글루탐산-Glu, 글루타민-Gln, 글리신-Gly, 히스티딘-His, 이소루신-Ile, 루신-Leu, 리신-Lys, 메티오닌-Met, 페닐알라닌-Phe, 프롤린-Pro, 세린-Ser, 트레오닌-Thr, 트립토판-Trp, 티로신-Tyr, 발린-Val.

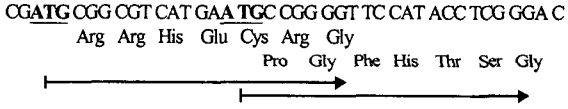

그림 1. DNA 번역 예

3. 면역계의 부정 선택 알고리즘

3.1 면역 시스템

생체의 방어체계인 면역계는 바이러스, 기생균, 병원균, 독소 등과 같은 항원이라고 통칭하는 매우 다양한 외부 유기체나 단백질에 대하여 생체를 방어할 수 있는 매우 정교하고 복잡한 시스템이다. 면역계를 구성하는 기본 요소는 두 가지 형태의 림프구이다. 이는 B세포(B 림프구)와 T세포(T 림프구)로써, B세포는 항체를 분비하는 체액성 반응을 하며, T세포는 면역에 관련된 세포를 자극 또는 억제하거나 감염된 세포를 죽이는 세포성 반응을 주로 담당한다[3, 4].

개체에는 각각 개인적인 특징을 나타내는 단백질이 존재한다. 이를 주조직 적합성 복합체(major histo-compatibility complex, MHC) 단백질이라 한다. T세포에는 MHC 단백질을 인식하는 부분이 존재하며 이를 이용해 자신의 세포여부를 판단하게 된다. 한편 B세포나 T세포는 특정 항원을 인식할 수 있는 인식부를 가지고 있으며 이를 항원 수용체(antigen receptor)라 한다[4].

여러 가지 면역세포 중 자기를 판별해주는 MHC 단백질 인식부와 항원을 인식하는 항원 수용체를 가지고 있는 세포는 T세포이다. T 세포는 이 두 가지 인식부를 가지고 항원에 의해 감염된 자기 세포를 인식한다. 따라서 면역계는 T 세포 생성시 MHC 단백질 인식부와 항원 수용체의 정상적인 동작여부를 확인하면서 T 세포를 생성한다. 이때 수용체의 정상적인 동작여부를 가리는 방법으로 사용되는 것이 긍정 선택(Positive Selection)과 부정 선택(Negative Selection)이다. 또한 B 세포도 항원 수용체를 생성하기 위해 부정 선택을 이용한다.

긍정 선택은 각 면역세포의 MHC 단백질 인식기능을 확인하는 선택방법이다. 자기세포에서 존재하는 MHC 단백질을 정확히 인지할 수 있는 면역세포만이 사용가능하기 때문에 갓 생성된 면역세포에 MHC 단백질을 결합시켜 긍정적인 선택이 되는 세포들로만 면역 세포를 구성한다. 이때 선택되지 않은 면역 세포들은 제거 된다.

부정 선택은 항원의 인식에 있어서 자기 세포를 항원으로 인식하는 것을 배제하기 위한 방법이다. 면역세포에 MHC 단백질을 결합시켰을 때 항원수용체가 MHC 단백질을 인식하지 못하는 세포들만 선택된다. 이때 긍정적인 선택을 하는 면역세포는 MHC 단백질을 항원으로 인식하는 세포들이므로 제거 된다.

그림 2는 생체 면역계에서 정상적인 면역 세포의 형성과정을 나타낸다. 그림의 면역세포는 긍정 선택과정을 거치면서 MHC 인식부가 결정되고 부정 선택과정을 거치면서 자신을 인식하지 않는 항원인식부가 결정된다. 따라서 두 가지 선택과정을 거친 T 세포는 자신과 항원을 인식할 수 있는 두개의 인식부를 가진다.

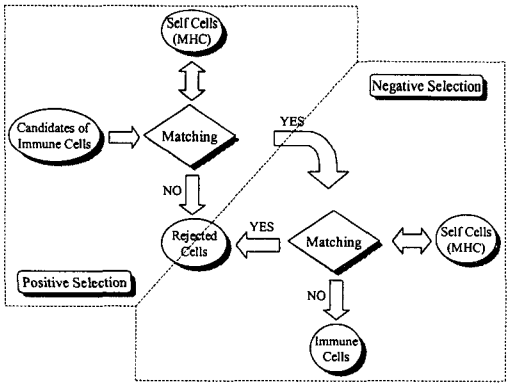

그림 2. T 세포의 형성과정

3.2 부정 선택 알고리즘

부정 선택에 기반한 변형 검지 알고리즘은 Forrest 등에 의해 제안된 자기-비자기 인식 알고리즘의 하나이다[5, 6]. 이는 자기 공간에 대해서 부정 선택을 거쳐 인식부 세트를 구성하고 이를 이용해 비자기 인식에 사용하였다. 이때 항원 인식부를 변형 인식부라 명명하였다. 변형 인식부(anomaly detector)를 이용한 비자기 인식 알고리즘은 자기 공간에 부분적으로 변경된 부분과 추가된 부분을 인식부를 통해 찾아낸다.

이 알고리즘은 크게 두 부분으로 구성된다. 하나는 자기 공간을 검사하기 위한 변형 인식부를 구성하는 부분이며 다른 하나는 구성된 변형 인식부를 이용하여 자기 공간을 모니터하며 변화의 발생을 검사하는 부분이다.

변형 인식부는 자기파일과 일치하지 않는 스트링(string)을 이용하여 구성한다. 즉 자기 파일과 일정한 길이 l의 랜덤 스트링 1bit씩 쉬프트(shift)하며 비교하여

매칭되지 않은 스트링을 변형 인식부로 구성한다. 이때 매칭은 q-인접 매칭 규칙을 사용한다.

이와 같은 과정을 통해 만들어진 변형 인식부들을 이용해 자기 파일이 아닌 부분을 인식한다. 이 알고리즘은 충분한 개수의 변형 인식부를 준비해 둠으로서 다양한 종류의 항원에 대해 인식할 수 있는 장점을 가지고 있다.

4. 부정 선택에 기반한 패턴 분류 알고리즘

4.1 염기에 기반한 방법

DNA는 4개의 염기로 구성되어있다. 따라서 변형인식 알고리즘의 이진 스트링 대신 4진의 DNA 스트링을 사용한다. 또한 매칭방식은 두 스트링의 해밍 거리(hamming distance)를 이용하여 임계값 M을 이용해 인식부를 선택한다. 인식부는 인식부와 패턴의 해밍 거리가 먼 것을 선택한다. 이때 인식부는 패턴을 한 염기씩 쉬프트하여 가장 작은 해밍거리 값을 이용한다.

$$HS(r, S) = min[H(r, s_i)] \quad (1)$$

$H()$는 해밍거리 r은 인식부, s_i는 패턴 S의 i번째 위치부터 인식부 길이만큼의 염기배열 이다. 그림 3은 부정 선택 방법에 의해 각 패턴에 대해 인식부 집합을 구성하는 방법을 나타낸다.

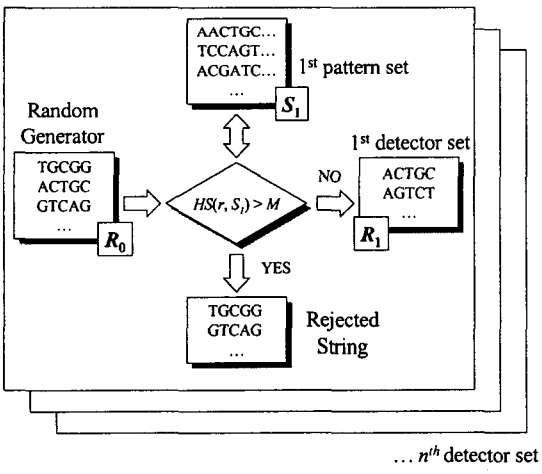

그림 3. 각 패턴에 따른 인식부 구성

이와 같은 과정에 의해 n개 패턴에 대한 n개의 인식부 집합이 생성된다. 이 인식부 집합을 이용해 새로운 입력 패턴에 대하여 어떤 패턴에 속하는지를 결정한다. 즉, 패턴인식 방법은 다음과 같다. 특정 패턴의 모든 인식부와 입력 패턴의 해밍 거리 즉 $HS(r, S)$가 M 이상이면 입력 패턴은 특정 패턴에 속한 것으로 결정한다.

4.2 아미노산에 기반한 방법

단백질의 군 분류나 구조 예측을 위해서는 DNA 염기 단위가 아닌 아미노산의 단위로 데이터를 처리하는 것이 필요하다. 따라서 DNA 염기 서열을 각 reading frame에 따라 아미노산으로 번역한 후 알고리즘을 적용한다. 이때 각 아미노산은 표 2와 같이 일련번호를 정하여 20개의 배열을 한 단위로 생각한다. 인식부 구성 및 패턴 인식 방법은 4.1절과 같다.

표 2. 아미노산과 순번

아미노산	Phe	Leu	Ile	Met	Val	Ser	Pro	Thr	Ala	Tyr
순번	1	2	3	4	5	6	7	8	9	10
아미노산	His	Gln	Asn	Lys	Asp	Glu	Cys	Trp	Arg	Gly
순번	11	12	13	14	15	16	17	18	19	20

5. 결 론

본 논문에서는 면역계의 부정 선택에 기반한 패턴 분류 알고리즘을 제안하였다. 부정 선택은 자기와 항원을 구별할 수 있는 면역세포의 항체를 생성하는 방법이다. 본 연구에서는 항체에 의한 자기 비자기 구별방법을 확장하여 n개의 패턴에 대하여 n개의 항원 인식부 셋을 구성함으로서 패턴 분류 하는 방법을 제안하였다. 신경망과 같은 기존의 패턴분류 방법은 패턴 인식을 위한 입력을 일정하게 유지시키는 것이 필요하다. 또한 패턴의 크기가 클수록 신경망의 크기가 커지게 된다. 따라서 특징을 추출을 통해 입력을 결정하여 사용한다. 하지만 제안한 방식은 *특징추출 과정이 필요 없으며 패턴의 크기가 매우 크고 크기가 일정하지 않을 때 매우 효과적*이다. 차후로 실험을 통하여 제안한 알고리즘의 유효성을 검증할 예정이다.

참 고 문 헌

[1] M. Gelfand, "Prediction of Function in DNA Sequence Analysis," Journal of Computational Biology, vol. 1, pp. 87-115, 1995.

[2] 김성동, 장병탁, "바이오 데이터 마이닝을 위한 기계학습 기법", 정보과학회지, 제18권, 제8호, pp. 63-72, 2000.

[3] R. A. Wallace, G. P. Sanders, and R. J. Ferl, BIOLOGY : The Science of Life, 3rd eds., HarperCollins Publishers Inc., 1991.

[4] I. Roitt, J. Brostoff, D. Male, Immunology, 4th edition, Mosby, 1996.

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

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