Backpropagation을 이용한 Promoter 예측 방법

허미영, 김홍기, 최진성
한국전자통신연구원

Prediction of Promoter by Backpropagation

Meeyoung Huh, Hongkee Kim, Jinsung Choi
Electronics and Telecommunications Research Institute
E-mail : hmy63069@etri.re.kr

Abstract

최근 생명공학 분야의 기술이 혁신적으로 발달함에 따라 게놈 프로젝트가 본래 계획보다 2 년 앞당겨져 2003 년 4 월 인간 유전자의 완전한 서열을 밝히고 성공적으로 완료됨으로서 관련 연구자들은 인간의 유전자에 대한 대량의 서열 데이터를 얻게 되었다. 그래서 게놈 프로젝트의 다음 단계로서 엄청난 양의 서열 정보 분석으로부터 유전자의 기능을 파악하고자 하는 연구들이 이미 세계적으로 활발히 진행되고 있다. 이러한 연구들의 최종적 목표는 질병 치료와 생명 연장의 실현이라고 볼 수 있다.

유전자 연구를 위해선 우선 일차적으로 유전자 부위를 파악해야 한다. 유전자는 구조적으로 다시 여러 부분으로 나뉘는데 유전자 발현의 개시에 매우 중요한 요소 중 하나가 바로 프로모터 (Promoter) 이다. 프로모터 내에는 TATA box 가 있는데 이는 프로모터의 핵심 요소이다. 프로모터는 생명체의 종 그리고 RNA 중합효소의 종류에 따라 다르다. 이 논문에서는 다양한 신경망 알고리즘 중의 하나인 Backpropagation 을 이용하여 밝혀지지 않은 서열에서 인간을 포함하는 원핵생물의 프로모터 서열을 예측할 수 있는 방법을 얻었기에 소개하고자 한다.

I. 서론

자연에 존재하는 생명체들이 그 생명을 유지하기 위해 생명체의 최소 단위라 할 수 있는 세포에서 많은 대사활동이 활발히 일어나는데 그 중에서 핵심이라 할 수 있는 과정이 유전자의 발현이다. 유전자가 RNA 로 전사되고 또 RNA 속에 담긴 암호를 해독해 단백질을 만드는 것이다. 최근 생명공학 분야에서는 수많은 미지의 유전자와 단백질의 기능을 밝히기 위해 노력하고 있다. 단백질을 생산하기 위해 유전자는 전사를 통해 messenger RNA (mRNA)를 만들게 되는데 이때 전사를 시작하기 위해 RNA 중합효소가 DNA 에 결합해야 하는데 그 결합 부위가 바로 프로모터이다. 유전자 발현에 관한 연구에서 있어서 Promoter 는 매우 중요한 요소이다.

RNA 중합효소는 세 종류로 RNA 중합효소 RNA 중합효소 Ⅱ, RNA 중합효소 가 있는데 각각 다른 유전자를 인식한다. RNA 중합효소 은 단백질을 생산하는 세포내 소기관인 리보솜의 구성요소인 ribosomal RNA (rRNA) 유전자에 관여한다. 또한 대부분의 단백질 관련 유전자와 RNA 프로세싱에 관여하는 것으로 알려진 small nuclear RNA 유전자는 RNA 중합효소 에 의해 전사가 시작된다. 그리고 RNA 중합효소 는 리보솜으로 유전자의 암호화된 코드에 맞게 아미노산을 전달하는 기능을 하는 transfer RNA (tRNA) 유전자의 발현 개시에 관여한다.

프로모터의 구조는 진핵생물과 원핵생물이 매우 다르며 원핵생물 중에서도 고등생물과 그 외 다른 생물들 간에 차이가 있다 (그림 1). 또한 RNA 중합효소의 종류에 따른 차이도 있는데 고등생물의 프로모터에서 RNA 중합효소 ,Ⅱ는 공통적으로 전사개시 위치로부터 25 base 떨어진 곳에 전사의

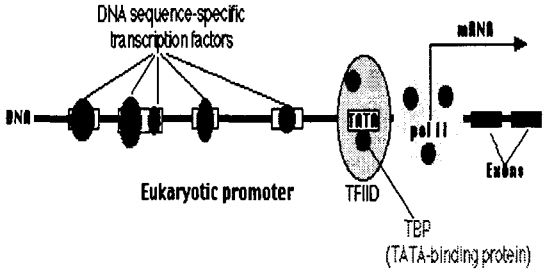

그림 1. 원핵생물의 프로모터 구조

핵심요소인 TATA box (consensus TATAWAW,W= A 또는 T)를 가진다. RNA 중합효소 는 -100 (전사개시 위치를 기준으로 전사방향의 반대쪽은 -, 전사방향은 +) 부근에 전사조절에 관여하는 Upstream control elements (UCE)가 있고 RNA 중합효소 Ⅱ는 전사 (-)방향으로 몇 백 base 떨어진 곳에 UCE 와 유사한 기능을 하는 여러 개의 다양한 Upstream elements 를 가지는데 이중에는 GC box (consensus GGGCGG) 또는 CCAAT box (consensus GCCAAT) 등이 있다.

본 논문에서는 신경망 이론 중에서도 여러 분야에서 매우 유용하게 이용되고 있는 Backpropagation 을 이용하여 원핵생물 프로모터의 핵심요소인 TATA box 가 포함된 주변서열을 이용하여 오직 아미노산 서열만 밝혀진 미지의 서열로부터 프로모터 부위를 예측함으로서 유전자 부위를 얻는데 도움이 될 수 있는 방법을 얻었기에 소개하고자 한다.

II. 실험방법

본 논문에서는 신경회로망 알고리즘 중의 하나인 Backpropagation 을 이용하였으며 학습에 사용한 재료는 사람의 유전자에서 대부분의 유전자 발현과 관련이 있는 RNA 중합효소 , Ⅱ의 프로모터 서열 중 핵심요소인 TATA box 를 포함하는 서열로서 -30 에서 -21 (10 잔기)까지의 서열과 프로모터 서열이 없는 사람의 케라틴 단백질 EST (Expressed sequence tag) 서열이다. 각각은 EPD (Eukaryotic Promoter Database)와 미국 국립생명공학 정보센터(NCBI)의 dbEST 로부터 추출하였다 (표 1, 2).

신경망의 학습을 위해 사람의 프로모터 서열과 케라틴 단백질 EST 는 각각 1 과 0 의 값으로 학습하였다.

구현한 방법의 테스트를 위하여 원핵생물 중에서 중복성이 없는 임의로 선택한 사람과 mouse 그리고 사람에 기생하여 사는 바이러스의 프로모터 서열을 이용하였으며 임의의 구간에서 10 잔기 길이의 서열을 추출하였다. 그리고 테스트에 사용된 사람 프로모터 서열은 학습에 이용된 서열과 중복되지 않는 데이터를 사용하였다 (표 3~5).

Gene name	Promoter sequence
snRNA	GACGGTGACG
Histone H1t	ATATAAGGCC
HMG-14	GGAGGGGGAG
TNP1	ATACCCAGAC
PRM2	CTTTATATAC
rp S14	TGACCCCCGT
ribosomal p. S19	CGACTTGTGC
keratin 67k	TCTCTATGCT
fibronectin	GTCCCATATA
P-glycoprotein 1	CTTCGCTCTC

표 1. 학습에 사용된 사람으로부터 얻은 임의의 중복성 없는 프로모터 서열.

dbEST Indenfier (GI)	Promoter sequence
26999372	TCTCGGACAC
26998814	TTTCATTGGC
20658968	TTGCAAAAGG
20495512	TTTTTTTTTT
20203307	CTAGTTGTAC
20203297	TGGGCGCCGG
20203164	CGCCGGCATC
20203157	TCATATTGGC
18086630	GGGGGCTACG
18086037	CCTGTCCTTT

표 2. 사람으로부터 얻은 케라틴 단백질의 est 서열.

Gene name	*Promoter sequence*
haptoglob Hp1F	GCATAAAAAG
2'5'-oligoA synt.	AGGAAACGAA
DHFR	CACAAATAGG
Thymidine kinase	GCCGTGGTTT
ADA	CCGTTAAGAA
TdT	ATCAAAACCC

TOP3	CTGTGGGGCG
ASS	TATAACCTGG
Arginase liver	TATAAATGGA
OAT	TTTAAGTTGC

표 3. 테스트에 사용된 사람으로부터 얻은 임의의 중복성 없는 프로모터 서열.

Gene name	Promoter sequence
Histone H1-C	ATATATACAG
Histone H2A.1 53	ATAAATGCTT
histone H2A 291A	ATATACTATG
hist. H2A(A)-613	ATAAAGGCCT
hist. H2A(B)-613	TAAATTCACC
histone H2B 221	TTAAAGAGCA
histone H2B 291A	AAATTGCAGC
histone H2B 291B	ATATAGGGCG
histone H2B-143	ATAAATAAAA
histone H3.1-A	TACTTAAAGG

표 4. 쥐로부터 얻은 임의의 중복성 없는 프로모터 서열.

Gene name	Promoter sequence
HCMV IE-I	CTATATAAGC
MCMV IE-I	GGTATAAGAG
HCMV IE gp/UL37	ATGTATATAA
HCMV IE gp/US3	CTATATATTC
HSV-1 IE-I	TGGGGTATAA
HSV-1 IE-II	GGTATAAGGA
HSV-1 IE-III	CTATATGAGC
HSV-2 IE-IV/V	TCGCGCACAT
HCMV b' DNA POL	TGTTTATAAT
HCMV b' 2.2 kb	ATGTATAAAT

표 5. 바이러스로부터 얻은 임의의 중복성 없는 프로모터 서열.

III. 결과 및 고찰

표 1 에서 사람의 프로모터와 케라틴 est 서열을 가지고 학습한 신경망을 임의로 추출한 쥐 프로모터 서열로 테스트한 결과를 보여준다 (기대값 1).

Promoter sequence	Results of Test
ATATATACAG	0.62
ATAAATGCTT	1.52
ATATACTATG	0.39
ATAAAGGCCT	1.00
TAAATTCACC	0.76
TTAAAGAGCA	1.01
AAATTGCAGC	0
ATATAGGGCG	0.90
ATAAATAAAA	2.00
TACTTAAAGG	0.47

표 6. 쥐의 프로모터 서열을 가지고 실행한 테스트의 결과값.

신경망의 학습 과정에서 1 과 0 의 값으로 학습하였음에도 불구하고 테스트 결과에서 이 범위를 넘는 값을 볼 수 있었다. 이는 학습과 표본추출 과정에서의 표본추출시 표본들간의 다양성의 부족, 즉 매우 다양한 유전자의 종류에 따른 프로모터 서열의 다양성을 충분히 반영하지 못한 것과 학습에 사용된 표본 수의 부족으로 이러한 결과가 나온 것으로 예측된다.

Promoter sequence	Results of Test
GCATAAAAAG	0.48
AGGAAACGAA	1.03
CACAAATAGG	0
GCCGTGGTTT	0.02
CCGTTAAGAA	0.44
ATCAAAACCC	0
CTGTGGGGCG	0.59
TATAACCTGG	1.34
TATAAATGGA	0
TTTAAGTTGC	0

표 7. 사람의 프로모터 서열을 가지고 실행한 테스트의 결과값.

쥐의 프로모터 서열을 가지고 실시한 테스트 결과 (표 5)보다 사람의 프로모터 서열을 가지고 실행한 테스트 (표 6)가 더 낮은 결과값을 보였다. 이는 사람 유전자의 다양성이 쥐 보다 더 높기 때문인 것으로 사료된다. 쥐의 경우에 1 을 넘는 오차값을 무시한다면 평균적으로 기댓값에 약 84 % 근접하였다.

Promoter sequence	Results of Test
CTATATAAGC	0.84
GGTATAAGAG	1.75
ATGTATATAA	0.95
CTATATATTC	0.69
TGGGGTATAA	1.77
GGTATAAGGA	1.55
CTATATGAGC	0.59
TCGCGCACAT	0.77
TGTTTATAAT	0.53
ATGTATAAAT	0.46

표 8. 바이러스의 프로모터 서열을 가지고 실행한 테스트의 결과값.

바이러스는 원핵생물에 포함되지만 사람과 진화적으로 거리가 먼 것으로 예상된다. 그러나 표 8 에서 바이러스의 프로모터 서열 중 임의로 선택한 서열로 테스트한 결과값은 쥐와 비슷한 결과를 보였다. 이는 실험에 사용된 바이러스가 사람에 기생하여 사는 바이러스이기 때문에 진화하는 과정에서 사람의 유전자 서열과 비슷한 서열을 운반하는 등의 사실과 관련되어 위와 같은 결과가 나온 것으로 사료된다.

IV. 결론

이미 밝혀진 프로모터 서열을 가지고 신경망을 학습시켜 실험에 사용되지 않은 서열에서 프로모터 부위의 존재 여부를 예측함으로서 미지의 서열로부터 프로모터를 예측할 수 있는 가능성을 보였다. 사람과 쥐 그리고 사람에 기생하는 바이러스의 프로모터 서열을 사용하여 구현된 방법을 평가 하였는데 생명체의 종 또는 조직 그리고 단백질의 종류가 다른 다양한 서열들로 각각 테스트 한다면 의미 있는 결과를 얻을 수 있을 것으로 사료된다.

V. 참고문헌

[1] V. Praz, R. Perier, C. Bonnard, P. Bucher, "The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data" , Nucleic Acids Research, vol. 30, pp. 322-324,2002.

[2] RC. Peier, V. Praz, T. Junier, C. Bonnard, P. Bucher, "The eukaryotic promoter database (EPD)" , Nucleic Acids Research, vol. 28, pp. 302-303,2000.

[3] W. Fujibuchi, M. Kanehisa, "Prediction of gene expression specificity by promoter sequence patterns", DNA Research, vol, 4, pp. 81-90, 1997.

[4] U. Ohler, GC. Liao, H. Niemann, GM. Rubin, "Computational analysis of core promoters in the Drosophila genome", Genome Biology , vol. 3, pp. RESEARCH0087-7, 2002.