PromSearch를 이용한 인간 염색체 22번의 프로모터 예측

김윤희° 김병희 장병탁
서울대학교 컴퓨터공학부 바이오지능 연구실
{yhkim98°, bhkim, btzhang}@bi.snu.ac.kr

Promoter Prediction on the Human Chromosome 22 by PromSearch

Yoonheui Kim°, Byoung-Hee Kim, and Byoung-Tak Zhang
Biointelligence Lab, Dept. of Computer Sci. & Eng., Seoul National University

요 약

PromSearch는 인간 DNA에서 코어 프로모터 영역을 예측하는 프로그램이며, PWM(position weight matrix)과 신경망을 기반으로 전사시작지점을 예측한다. 프로그램은 대량의 서열 데이터를 처리할 수 있도록 구성되었으며, 본 논문에서는 인간 염색체 22번에 대한 프로모터 예측 결과를 제시한다. Annotated된 936개의 유전자와 PromSearch가 예측한 프로모터 간의 위치의 상관관계를 계산한 결과 87개에 대해 프로모터 예측 결과가 의미 있는 것으로 밝혀졌다. 예측의 민감도는 25%이며, PromSearch가 대규모 시퀀싱 프로젝트에서 나오는 대량의 서열 데이터를 1차적으로 분석하는 도구로서 사용될 수 있음을 확인하였다.

1. 서 론

염기서열을 대량으로 빠르게 분석하는 기술이 발전함에 따라, 공용 도메인에 수많은 염기 서열이 쏟아지고 있다. 실험적인 분석만으로는 염색체 수준에서 유전자를 탐색하고 이들 간의 조절 관계를 파악하는 것은 시간과 노력이 많이 필요하다. 이에 따라 컴퓨터를 이용한 빠르고 정확한 주석(annotation) 작업의 중요성이 증가하고 있다.

프로모터는 DNA 상에서 유전자의 발현을 조절하는 영역이며, 프로모터 위치를 파악하는 일은 새로운 유전자를 찾아내고, 그 기능을 밝히며 나아가 유전자 간의 관계를 규명하기 위한 근간이 되는 문제이다.

1990 년대에 많은 프로모터 예측 프로그램이 나왔으나, 대부분이 좁은 영역만을 처리하고 무엇보다도 false positive 가 높다는 점이 큰 문제였다. 최근 PromoterInspector[1], DPF[2] 등 대규모의 서열을 처리할 수 있고 성능이 크게 개선된 프로그램이 소개되어, 게놈 수준에서의 유전자 조절 작용을 분석하는 데 일조를 할 수 있게 되었다.

본 논문에서는 인간 염색체 22 번에 대한 프로모터 예측 결과를 토대로 PromSearch 가 전체 게놈에 대한 초기 주석(annotation)에 사용할 수 있는 적절한 도구임을 제시하고자 한다.

2. 실험 도구 및 데이터

2.1 PromSearch

PromSearch[3]는 DNA 서열을 슬라이딩 윈도우(sliding window) 방식으로 검색하여 코어 프로모터의 전사시작지점(TSS)을 예측한다. 윈도우의 크기는 300bp 를 기준으로 하였으며, 이 윈도우에 대한 분할 모델을 설정하였다. 즉, 윈도우를 네 개의 프로모터 요소(TATA box, Inr, GC box, CAAT box)와 DPE(downstream promoter element) 및 인접(proximal) 프로모터의 영역으로 분할하고, 각 영역에 대해 [4]의 PWM(position weight matrix) 및 DPF(dragon promoter finder)[2]의 펜타머(pentamer)의 PWM 을 이용하여 프로파일을 생성한다. 프로파일의 수는 10 개이며 이 프로파일의 조합으로 프로모터를 예측할 분류기(classifier)로 다층 퍼셉트론(multi-layer perceptron)을 사용하며, 성능의 향상을 위해 AdaBoost.M1 알고리즘을 신경망에 적용하였다. [5]의 데이터에 대한 평가 결과 민감도(sensitivity) 41.7%, 특수도(specificity)는 27FP, 1/1230bp 로 나타났다. Fickett 데이터를 포함하여 여러데이터에 대해 NNPP 2.1 과 Promoter 2.0 보다 높은

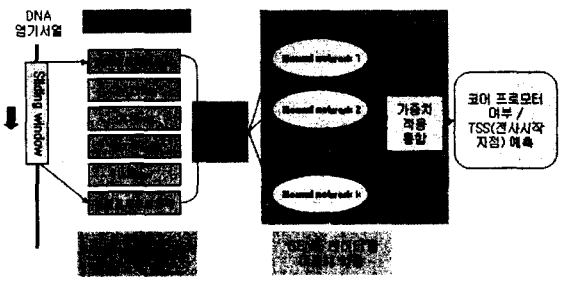

그림 1. PromSearch 동작 알고리즘 구성

성능을, TSSG 와는 유사한 성능을 보였다.

2.2 인간 염색체 22번

인간 염색체 22 번은 인간의 유전자 정보를 밝혀내는 인간게놈프로젝트의 성과로써 거의 완벽하게 분석되어 발표된 첫 번째 결과물이다. [6]이 발표된 이후로 인간염색체 22 번에 대해 계속적인 분석 및 주석 달기(annotation)가 계속되었고, 2004 년 현재까지 20 개 이상의 프로모터와 936 개에 달하는 유전자의 시작점 위치가 밝혀졌다. 인간 염색체 22 번은 완성된 하나의 염기서열로 PromSearch 가 목표로 하는 23(22 의 상염색체와 X, Y 염색체)개의 염색체 중 하나로써 이 염색체의 염기서열을 PromSearch 로 분석한 결과는 PromSearch 의 성능을 보여주는 객관적인 지표가 될 수 있다.

3. 실험 및 결과

3.1 실험 방법

본 논문의 실험은 바이오 지능 연구실에서 개발한 인간 프로모터의 위치를 탐색하는 프로그램 PromSearch 를 이용하여 진행되었다.[3] PromSearch 는 유전자 서열만으로 프로모터의 위치를 탐색하는 프로그램으로서 인간 염색체 22 번 서열에 대해 두 가지 설정을 적용하여 분석하였다. 두 설정을 적용한 결과 각각 22 번 염색체 전체에서 1881, 5335 개의 후보 TSS 를 얻었다. 예측결과 평가를 위하여 실험적으로 검증된 프로모터 서열들([7]에서 사용된 프로모터 서열들)과 웹에 발표된 주석달린 유전자 시작점(annotated gene start)들을 PromSearch 로 예측한 결과와 비교하는데 사용하였다 (http://www.sanger.ac.uk/HGP/chr22). 특히 주석달린 유전자 시작점들과 PromSearch 결과의 상관관계를 계산하는 GenomeInspector 프로그램을 이용하였다. (http://www.gsf.de/biodv/genomeinspector.html) GenomeInspector 는 수 메가 바이트에 달하는 염기서열 상의 구간들과 점들 혹은 점들과 점들의 거리의 상관관계를 쉽게 구할 수 있다. 여기서는 PromSearch 로 예측된 결과 TSS 의 하부(downstream)로 수 bp 떨어진 거리 내의 주석달린 유전자 시작점들의 개수를 측정하거나, 예측된 프로모터 구간들 근방으로 주석달린 유전자 시작점들이 집중되어 있음을 보여 PromSearch 의 성능을 보이는데 쓰였다.

3.2 실험적으로 알려진 프로모터 예측 결과

22번 염색체 전체에 대한 결과 분석에 앞서 개략적인 성능 평가를 위해 [7]에 수록된 실험적으로 알려진 22번 염색체 상의 프로모터 중 몇 개를 정확히 예측하는지 살펴보았다. [7]에는 해당 유전자의 주석달린 유전자 시작점(annotated gene start)과 프로모터 서열간의 상대적인 위치가 주어져 있는데 이를 이용해 염색체 22번의 전체서열에서 해당 유전자의 프로모터 서열의 위치를 계산하였다. 이 계산된 프로모터의 구간안에 PromSearch가 예측한 전사시작지점이 놓이면 그 프로모터를 예측한 것으로 보았을 때, 예측결과는 표 1과 같다.

구분	PromSearch로 예측된 알려진 프로모터의 개수(해당 유전자)	비율 (%)
설정 1	3 (HMOX1, LIF, LIMK2)	15
설정 2	5 (HMOX1, GGT1, EWSR1, LIF, LIMK2)	25

표 1. 실험적으로 알려진 20개 프로모터에 대한 예측 결과

3.3 주석달린 유전자 시작점(annotated gene start)을 기준으로 한 분석

본 논문에서는 주석달린 유전자 시작점을 이용하여 PromSearch의 프로모터 예측 결과의 정확도를 분석하는데 적용하였다. 프로모터는 유전자 발현을 조절하므로 많은 경우 유전자 시작점에 가까이 위치한다. 그 예로 [7]에서 사용된 실험적으로 알려진 프로모터의 위치는 대부분 주석달린 유전자 시작점에 가까이 위치하고 있다. 따라서 예측된 전사시작지점(TSS)를 기준으로 하부(downstream)로 555bp 안에 주석달린 유전자 시작점이 위치하면 예측이 정확한 것으로 하여 측정한 결과가 표2 이다. 여기서 555bp라는 수치는 [7]에서 예측된 프로모터 영역의 평균 길이이며, 이 값을 사용하여 [7]논문과의 결과를 대등한 조건에서 비교할 수 있도록 하였다. 그림 2는 예측된 TSS의 상부(upstream)로 250bp, 하부(downstream)로 50bp가 되는 구간들로부터 주석달린 유전자 시작점까지의 거리에 따른 유전자 시작점의 분포를 보여준다. -6000bp부터 3000bp까지의 회색영역에 가유전자(pseudo gene)를 제외한 총 유전자 700개 중 385개(55%), 코딩 유전자 393개 중 194개(49%)가 위치하여 주석달린 유전자 시작점들이 프로모터 구간 근처에 집중적으로 분포하고 있음을 보여준다.

유전자 그룹	#	예측된 TSS의 수가 1881개일 때			예측된 TSS의 수가 5355개일 때		
		절대값	상대값(%)	O값	절대값	상대값(%)	O값
모든 유전자	930	31	3.3	4.2	87	9.4	4.3
가유전자(pseudo gene)를 제외한 유전자	700	24	3.4	4.8	66	9.4	5.2
가유전자 (pseudo gene)	230	7	3.0	1.3	21	9.1	2.5
추가적인 예측		1850			5268		

표 2. 인간 염색체 22번에서의 PromSearch의 두 예측과 주석달린 유전자 시작점과의 관계

: 주석달린 유전자 시작점(annotated gene start)의 갯수
절대값 : PromSearch로 예측된 위치로부터 정해진 구간(555bp)안에 있는 주석달린 유전자 시작점의 갯수
상대값 : 전체 주석달린 유전자 시작점에 대한 PromSearch로 예측된 주석달린 유전자 시작점의 비율
O값 : 바르게 예측된 위치가 고르게 분포할 때 발견되었을 유전자 시작점 갯수에 대한 바르게 예측된 시작점의 개수의 비율

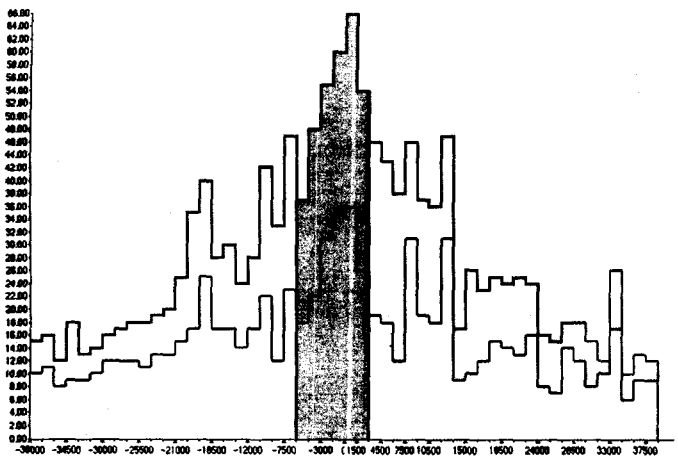

그림 2. 인간 염색체 22번에서 PromSearch 로 예측한 프로모터 구간과 주석달린 유전자 시작점(annotated gene start)과의 상관관계 분석. 세로축은 PromSearch에서 예측된 프로모터 구간과 상대적인 거리 안에 있는 주석달린 유전자 시작점의 갯수를, 가로축은 PromSearch에서 예측된 프로모터와 주석달린 유전자 시작점과의 상대적인 거리를 나타낸다. 가로축 영점을 기준으로 +값은 프로모터 구간을 기준으로 주석달린 유전자 시작점이 하부(downstream) 방향에 위치했다는 의미이고 -값은 반대로 상부(upstream)쪽으로 위치한 것을 나타낸다. 위쪽의 그래프는 가유전자(pseudo gene)을 제외한 모든 유전자에 대한 상관관계를 보여주고 아래쪽의 그래프는 주석달린 유전자 시작점 중 코딩 유전자만에 대한 상관관계 그래프이다. 즉 위의 그래프는 아래 그래프의 결과값을 모두 포함한다. 중앙의 회색 영역(-6000bp~3000bp)에 가유전자를 제외한 유전자의 55%가 집중되어 있어 예측된 TSS주변(가로축 상 영점)으로 주석달린 전사 시작점이 집중되어 있음을 보여준다.

4. 결론 및 논의

[7]에 따르면 PromoterInspector는 2001년 당시 알려져 있던 22번 염색체의 545개의 *annotated gene*에 의해 확인할 수 있는 180개의 프로모터 영역을 예측하였으며, 대략 3개 중 하나의 비율로 정확한 예측을 하였다. PromSearch는 930개 중에서 87개를 예측한 반면, 60개 중 하나의 비율로 정확한 예측을 하여 Promoter-Inspector와 같은 " chromosal scaffolding" [7], 즉 긴 서열에서 유전자를 기준으로 영역을 설정하여 탐색 공간을 줄이는 역할을 하기에는 부족하다.

그러나, PromSearch는 표 1에서 보였듯이, 유전자의 시작점을 알고 있는 상태에서 25%의 민감도(sensitivity)를 보였으며, 그림 2의 결과로 보아 볼 수 있듯이 유전자의 시작점을 안다면 이를 중심으로 9000bp내의 범위에서 55% 정도의 프로모터를 찾아낼 수 있다.

또한, Pentium4 1G, 256M RAM, MS windows XP환경에서 분당 2M bp 정도의 서열을 처리하였다.

정리하면, PromSearch 는 대량의 서열을 1 차적으로 빠르게 처리하여 프로모터를 탐색하기 위한 공간을 줄여주고, 유전자의 위치를 아는 상태에서는 비교적 신뢰할 수 있는 프로모터 예측을 한다고 할 수 있다.

감사의 글

본 연구는 과학기술부 국가 지정연구실사업에 의하여 일부 지원되었음

참고문헌

[1] M. Scherf, A. Klingenhoff, and T. Werner, Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol.*, 297:599-606, 2000.

[2] V. B. Bajic, et al., Computer model for recognition of functional transcription start sites in polymerase II promoters of vertebrates, *Journal of Molecular Graphics & Modeling*, 21(5): 323-332, 2003.

[3] 김병희, 김윤희, 남진우, 임명은, 심정섭, 박선희 장병탁, PromSearch: 신경망을 이용한 코어 프로모터 예측 프로그램, *제30회 한국정보과학회 가을 학술발표 논문집 (II)*, 제30권 2호, pp. 769-771, 2003.

[4] P. Bucher, Weight matrix description of four eukaryotic RNA polymerase II promotor elements derived from 502 unrelated promotor sequences. J Mol Biol., 212:563-578, 1990.

[5] J. W. Fickett and A. G. Hatzigeorgiou, Eukaryotic promoter recognition. *Genome Res.*, 7:861-878, 1997.

[6] I. Dunham, et al., The DNA sequence of human chromosome 22. *Nature*, 402:489-495.

[7] M. Scherf, et al., First pass annotation of promoters on human chromosome 22, *Genome Res.*, 11:333-340, 2001.