Intelligent System for Promoter Recognition with Multiple Decision Models

프로모터 예측을 위한 다중 결정 모델 지능 시스템

Sang-Soo Yeo[1], Jung-Won Rhee[1], Sung-Kwon Kim[1*]

[1] School of Computer Science & Engineering, Chung-Ang University, Seoul, Korea

*To whom correspondence should be addressed. E-mail: skkim@cau.ac.kr

Abstract

The Development of promoter recognition systems is a interesting problem in computational biology. In this paper, we introduce a intelligent system for promoter recognition with multiple decision models using artificial neural networks. We have trained this models with 1871 human promoter sequences and 5230 exon and intron sequences. Our system is found to perform better than other promoter finding systems in sensitivity and specificity measures. We have tested our system with Chromosome 22 dataset.

Introduction

프로모터는 일반적으로 DNA Sequence 상에서 유전자 발현을 위해서 RNA Polymerase II와 다양한 TF(Transcription Factor)들이 결합하는 부분을 일컫는 용어이다. 일반적으로 전사가 시작되는 TSS (Transcription Start Site)를 주위로 -500~+100bps정도의 범위를 프로모터라고 본다. 진핵 생물(eukaryotic)에서는 프로모터가 종별로 잘 보존되어 있지 않고, 그 패턴이 일정하지 않다. 따라서, 일반적인 패턴 매칭 방법으로 프로모터를 예측하는 것은 좋은 결과를 얻어내기 힘든 것으로 알려져 있다[1].

본 연구는 보건복지부 IMT-2000 출연금 기술개발사업의 지원에 의하여 이루어진 것임. (01-PJ11-PG9-01BT00B-0020)

생물학적으로 전사의 과정에 대한 이해가 완벽하지 못한 상태이기 때문에 프로모터를 예측하는 알고리즘에 대한 연구는 매우 가치 있는 연구라고 할 수 있다. 여러 연구를 통해서 진핵 생물 또는 척추 동물의 프로모터를 예측하는 시스템들이 개발되었다. 1997년까지의 연구는 Fickett의 논문[2]에 잘 비교 정리 되어 있다. NNPP2.1[3], Promoter2.0 [4], Promoter Insptecor[5], Dragon Promoter Finder[6] 등이 현재 많이 알려진 프로모터 예측 시스템들이다.

앞서 말한 바와 같이 진핵 생물에서 프로모터의 시퀀스 패턴은 일정치 않고 다양하기 때문에 현재까지의 프로그램들은 일정한 수준의 True Positive(TP) 예측을 해내기 위해서는 False Positive(FP) 예측이 매우 많이 나

오게 된다[2]. 위의 시스템들도 민감도와 특이도라는 평가 기준을 통해서 프로모터 예측의 성능을 설명하였다.

민감도와 특이도는 예측의 결과가 실제의 결과에 얼마나 부합되는지를 알려주는 평가 기준들이다. 아래에 정의되어 있다.

$$\text{민감도}(sensitivity) = \frac{TP}{TP+FN}$$

$$\text{특이도}(specificity) = \frac{TP}{TP+FP}$$

여기서 TP는 프로모터로 예측된 값이 실제로 프로모터인 경우를 말하고, FN은 예측은 프로모터가 아닌 것으로 나왔지만 실제로는 프로모터인 경우를 말한다. FP는 예측은 프로모터라고 예측되었지만, 실제로는 프로모터가 아닌 경우를 말한다.

민감도는 실제 프로모터의 총 수에 대한 예측된 프로모터의 수를 말하며 높을수록 좋다. 특이도는 예측된 프로모터 중에서 실제로는 프로모터가 아닌 것의 수를 말한다. 특이도 역시 높을수록 좋은 결과를 의미한다.

본 논문에서 개발된 시스템은 인간 프로모터 예측을 위해 만들어졌고, 평균적으로 기존의 프로모터 예측 시스템에 비해서 민감도(sensitivity)가 많이 향상된 결과를 보여준다.

Methods

본 시스템은 Dragon Promoter Finder와 비슷하게 3개의 독립적인 프로모터 예측 모델를 가지고 있다[6]. 각각의 모델들은 다양한 실험을 통해서 미리 정의된 파라메터들에 의해 민감도와 특이도 레벨이 다르게 설정되어져 있다. 그러나, 3개의 모델 모두 기본 구조는 동일하다.

예측 모델의 기본 구조는 지금까지의 시스템들과 달리 매우 독특한 접근 방식을 취했다. 각 모델은 프로모터 데이터들과 비프로모터 데이터들을 가지고 인공지능 신경망(ANN, artificial neural network)을 학습 시켜 만들어진 결정 모델(decision model)이다. 일반적인 결정 모델은 예' 또는 아니오'의 답을 하도록 학습 시키는 것이 일반적이다. 그러나, 본 시스템에서는 학습을 하기 이전에 프로모터 데이터들을 먼저 클러스터링 과정을 통해 10개의 클러스터들로 나눈다. 마찬가지로 비프로모터 데이터들도 10개의 클러스터들로 나눈다. 그런후 신경망 학습 과정을 통해서 20개 중에서 하나(또는 그 이상)의 답을 하는 결정 모델을 만든다.

프로모터 데이터들을 클러스터링 해서 여러 개의 클러스터들로 만들게 되면, 다양한 종류의 프로모터에 대해서 민감도를 높일 수 있는 가능성이 그만큼 많아진다고 볼 수 있다. 이는 비프로모터 데이터들에 대해서도 마찬가지이다. 한 가지 형태의 비프로모터 데이터가 아니라 여러 개의 클러스터들로 구분되어진 비프로모터 데이터들은 특이도를 높일 수 있는 가능성이 높아진다고 볼 수 있다. 비슷한 예로 Dragon Promoter Finder[6]에서는 데이터들을 프로모터, 엑손, 인트론으로 나누어서 학습을 시켰다. 본 시스템에서 프로모터와 비프로모터를 각각 10개의 클러스터로 나눈 것은 특별한 이유가 있는 것은 아니며 여러 번의 시도를 통해서 선택된 적절한 클러스터의 개수이다.

Datasets

Training Datasets

본 시스템의 학습을 위해서 사용한 데이터들은 는 다음과 같다. 프로모터 데이터들

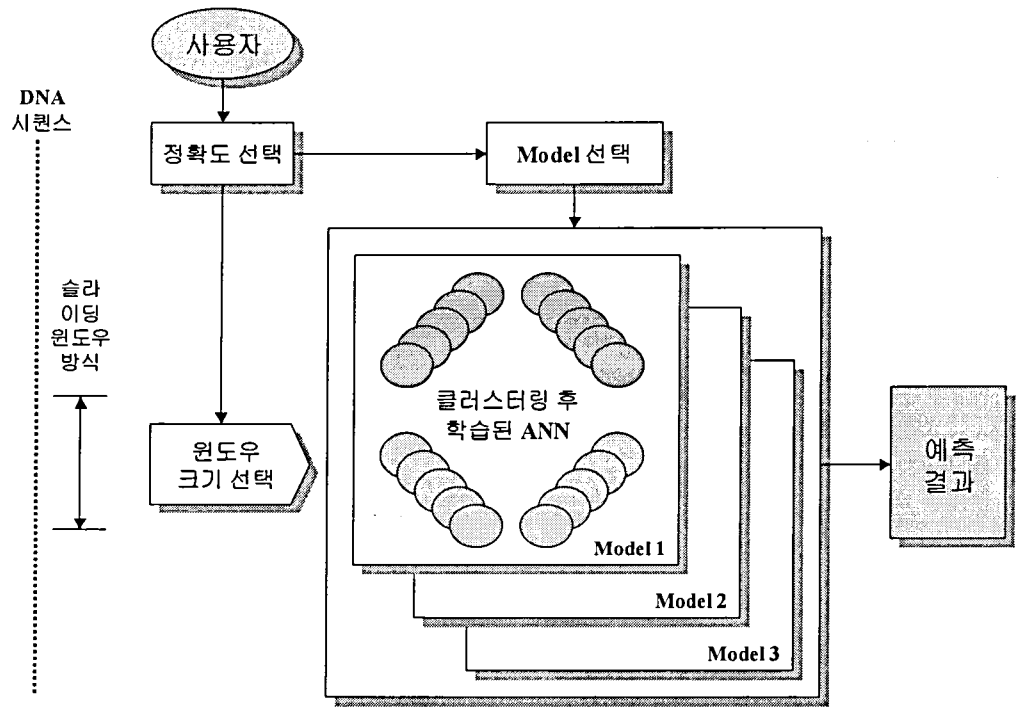

그림 1 개발된 시스템의 구성도

은 EPD(Eukaryotic Promoter Database. rel. 76)로부터 가져왔다. EPD에는 현재 2997개의 진핵 생물의 프로모터 시퀀스 데이터가 있는데, 이중에서 인간(Homo Sapiens) 프로모터 데이터 1871개를 본 시스템의 프로모터 학습 데이터로 사용하였다. 프로모터 시퀀스는 TSS를 중심으로 -200bps ~ +50bps를 취하였다. 따라서, 프로모터 시퀀스의 길이는 251bps이다.

비프로모터 데이터들은 Genbank로부터 가져왔다. 먼저 EPD에서 가져온 프로모터 시퀀스의 뒤쪽에 위치하는 엑손과 인트론들의 시퀀스들을 가지고 와서 중복된 부분들을 없애는 과정을 거치고 251bps씩의 길이로 잘라서 3934개의 시퀀스를 만들었고, 인트론만으로 이루어진 1296개의 중복되지않는 시퀀스를 역시 GenBank로부터 획득하였다.

학습 데이터들을 가지고 신경망을 학습시키는 과정에서 20% Cross Validation을 하였다.

Evaluation Datasets

본 시스템의 프로모터 예측 성능을 평가하기 위해서 사용된 테스트 데이터는 다음과 같다. 첫번째 데이터 그룹은 다른 유전자 예측 프로그램들에서 학습(training) 데이터로 사용된 데이터들로써, GeneId[7] 시스템(http://www1.imim.es/geneid.html), Genie[8] 시스템(http://www.fruitfly.org/seq_tools/datasets/Human) 등에서 획득하였다.

두 번째 데이터 그룹은 염색체 22번 시퀀스이다. 염색체 22번 데이터는 Sanger 센터(http://www.sanger.ac.uk/HGP/Chr22/)에서 만들어진 시퀀스 데이터이다. 염색체 22번은 총 47748585bps이고, 1~13100000까지에는 유전

자가 존재하지 않는 것으로 밝혀진 부분이다. 13100001~47748585에는 현재 936개의 유전자가 밝혀져 있다. 여기에는 coding genes, partial genes, non-coding RNA genes, pseudogenes, IGLV/J(Immunoglobulin joining and variable regions) 등이 포함되어 있다.

Results

본 시스템은 다른 프로모터 예측 시스템에 비해 민감도에서 많은 향상이 있었다. 특이도 또한 많이 알려진 Dragon Promoter Finder v1.2에 비해 향상된 결과를 보여준다. 다음의 표는 본 시스템과 다른 프로모터 예측 시스템의 민감도와 특이도를 비교한 표이다.

Table 2 인간염색체 22번에 대한 실험 결과 비교

개발된 시스템		Dragon Promoter Finder		Promoter Inspector	
민감도	FP/TP	민감도	FP/TP	민감도	FP/TP
91%	7.33	60.177%	3.64	45%	1.975
63%	3.45	30.67%	2.06		
56%	2.33	20.06%	1.0		

Discussion

본 시스템의 다중 결정 모델을 가지고 프로모터를 예측하는 시스템으로서 민감도에서 많은 향상이 있었고, 그에 대한 특이도에서도 다른 프로모터 예측 시스템과 비교했을 때 향상된 결과를 보여준다.

특이도면에서의 성능향상은 계속해서 연구되어야 하는 것으로서, 연구 모델의 변경, 파라메터 값의 조정 등이 더 필요하다고 본다.

References

[1] A.G. Pederson, P. Baldi, Y. Chauvin and S. Brunak, The biology of eukaryotic promoter prediction a review, *Comput. Chem.*, **23**, 191-207, 1999

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

[3] M.G. Reese, N.L. Harris and F.H. Eeckman, Large scale sequencing specific neural networks for promoter and splice site recognition. *In Biocomputing Proceedings of the 1996 Pacific Symposium*, 1996

[4] S. Knudsen, Promoter2.0: for recognition of Pol II promoter sequences, *Bioinformatics*, **15**, 356-361, 1999

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

[6] V.B. Bajic, S.H. Seah, A. Chong, G. Zhang, J.L.Y. Koh and V. Brusic, Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters, *Bioinformatics*, **18**, 198-199, 2002

[7] R. Guigo, S. Knudsen, N. Drake and T. Smith, Prediction of Gene Structure, *J. Mol. Biol.*, **226**,141-157, 1992

[8] M. Reese, D. Kulp, A. Gentles and U. Ohler, http://www.fruitfly.org/seq_tools/datasets /Human, 1999