PromSearch: 신경망을 이용한 코어 프로모터 예측 프로그램

김병희[O,1], 김윤희[1], 남진우[1], 임명은[2], 심정섭[2], 박선희[2], 장병탁[1]
[1]서울대학교 컴퓨터공학부 바이오지능연구실, [2]한국전자통신연구원
[1]{bhkim[O], yhkim98, jwnam}@bi.snu.ac.kr,
[2]{melim, simjs, shp}@etri.re.kr, [1]btzhang@cse.snu.ac.kr

PromSearch: a core-promoter prediction program using neural networks

Byoung-Hee Kim[O,1], Yun-Hee Kim[1], Jin-Woo Nam[1], Myung-Eun Lim[2], Jeong-Seob Sim[2], Sun-Hee Park[2], and Byoung-Tak Zhang[1]
[1]Biointelligence Lab, School of Computer Sci. & Eng., Seoul National University
[2]Electronics and Telecommunications Research Institute

요 약

PromSearch는 DNA 염기서열 상에서 프로모터의 위치를 예측하는 프로그램이다. 다루는 대상은 인간 DNA의 프로모터이며, 프로모터의 TSS(transcription start site, 전사시작지점)를 예측하는 것을 목표로 한다. 프로모터 영역을 세분하여 각 영역에 대한 프로파일을 PWM(position weight matrix)을 이용해 작성하며, 임의의 염기서열이 입력으로 주어지면 세분한 영역의 점수를 신경망을 이용해 통합하여 프로모터 여부와 TSS의 위치를 결정한다. 프로모터 영역의 분할은 코어 프로모터의 구성 요소인 TATA-box와 Inr, DPE(downstream promoter element), 그리고 코어 프로모터의 위쪽으로 150bp 크기의 영역 등으로 4분할하였다. Fickett의 데이터를 이용한 평가 결과 sensitivity 43%, specificity 88fp(1/376bp)의 성능을 보였다.

1. 서 론

프로모터(promoter)란 유전자(gene)가 언제 어디서 어느 정도 발현할 것인가를 결정하는 작용을 하는 염기 서열로서, 지령 기능을 갖는 서열이라고 할 수 있다. 유전자가 발현하기 위해서는, 유전자의 앞쪽에 존재하는 프로모터 영역에 다양한 단백질이 결합하여야 한다. 진핵세포(eukaryote)의 경우에는 그 중에서도 RNA중합효소(RNA polymerase)라고 하는 단백질이 코어 프로모터 영역에 붙는 과정이 반드시 필요하므로 일반적으로 진핵세포의 프로모터는 <그림 1>에서와 같이 'Polymerase II' 또는 간략히 Pol-II프로모터라고 불린다. Pol-II 프로모터는 DNA에서 RNA로 서열이 복사되는 '전사'가 시작되는 '전사시작지점(TSS, transcription start site)'을 중심으로 그 위치에 따라 <그림 1>과 같이 여러 영역으로 구분할 수 있다. 각 영역에는 전사를 조절하는 데 관여하는 단백질(TF, transcription factor)들이 결합하는 부위가 모여 있으며, 그 분포와 조합은 매우 다양하고 복잡하다. 이런 점에서 in silico로 프로모터를 예측하는 일은 쉬운 일이 아니며, 지금까지 개발된 많은 프로모터 예측 시스템들은 FP(false positive)가 높아서 아직 실용적인 수준의 성능을 보이지 못하고 있다.

본 논문에서는 PromSearch라 명명한 프로모터 예측 프로그램의 알고리즘과 성능 평가 결과를 소개한다.

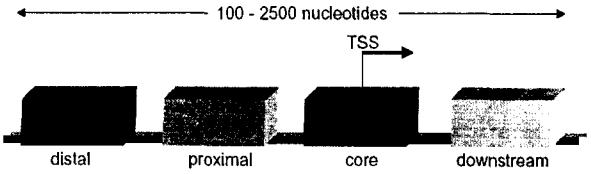

그림 1. Polymerase II 프로모터의 구성

PromSearch는 코어 프로모터를 포함한 프로모터 영역을 분할하여 각 영역에 대한 프로파일을 PWM(position weight matrix)을 이용하여 생성한 후 이를 신경망을 이용해 통합하여 프로모터 여부를 결정한다.

2. 알고리즘

2.1 관련 연구

지금까지 Pol-II프로모터 예측을 위해 다양한 접근방법이 시도되어왔다. 이들 방법들은 크게 다양한 프로모터에 일반적으로 적용되는 모델 설정, 특정 프로모터에 초점을 둔 모델링 및 여러 종간의 유사성(homology)에 기반을 둔 접근법으로 분류할 수 있다. 현재까지 가장 많은 관심과 결과가 나온 접근법은, 일반적인 프로모터를 예측하기 위한 모델링이다. 이러한 접근법은 다시 내용에 의한 탐색(search-by-content)과 신호에 의한 탐색(search-by-signal) 및 둘을 조합한 방법으로 구분할 수 있다 [1].

코어 프로모터에 대한 비교적 정확한 모델들 중 많은 수가 신경망(neural network)을 기반으로 하고 있으며, DPF [2], NNPP[3], Promoter2.0[4], McPromoter[1] 등이 대표적인 예다.

2.2 PromSearch의 구조

2.2.1 프로모터 모델링

PromSearch의 기반 알고리즘은 앞 절에서 기술한 접근방법 중 '신호에 의한 탐색'으로 분류할 수 있는 특성을 가진다. PromSearch에서 사용된 프로모터 모델은 <그림 2>와 같다. 많은 인간 DNA상의 프로모터에서 공통적으로 발견되는 두 코어 프로모터 요소, TATA-box와 Inr영역의 위치는 가변적이며, 이를 반영하기 위해 본 모델에서는 두 요소의 PWM의 적용 범위를 유동적으로 설정한다.

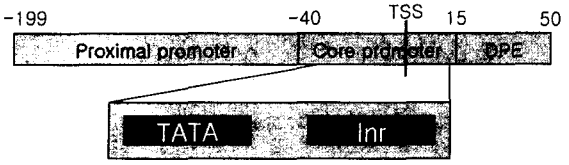

그림 2. PromSearch에 사용된 프로모터 모델. 코어 프로모터의 두 요소 (TATA-box, Inr) 및 3'쪽의 DPE, 5'쪽의 영역 등 4개의 세그먼트로 구분

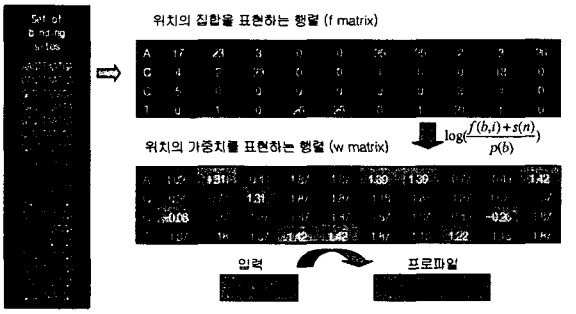

그림 3. PWM(position-weight matrix) 생성 및 서열의 프로파일링

2.2.2 PWM

PWM은 DNA 서열 분석에서 각 염기의 분포 모델링과 서열의 프로파일링에 가장 많이 응용되는 도구 중의 하나이며 <그림 3>과 같은 과정을 거쳐 생성한다. Bucher의 PatOp 알고리즘[5]을 이용해 생성한 네 가지 프로모터 구성 요소 (TATA-box, Inr, CAAT-box, GC-box)의 PWM은 프로모터 분석과 예측에 중요한 도구로 사용되어 왔다. 최근에 서비스를 시작한 SSA(signal search analysis)[6, http://www.isrec.isb-sib.ch/ssa/]를 통해 지난 10여년간 추가된 데이터를 반영한 새로운 PWM을 생성할 수 있게 되었다.

본 연구에서는 SSA를 이용해 EPD(eukaryotic promoter database)의 척추동물 프로모터 집합에서 생성한 TATA-box 및 Inr의 PWM을 사용하였다. 최적화 과정을 거쳐 생성된 이 PWM에는 프로파일의 중요도를 결정하는 임계값으로서의 cut-off value 및, PWM을 적용할 수 있는 범위가 부수적으로 주어진다.

TATA PWM은 [-40, 14], Inr PWM은 [-7, 14] 영역 내에서 유동적으로 적용되며, 해당 영역에서의 최고값을 프로파일로 사용한다. 또한, Inr PWM에 대해 최고값을 가지는 영역의 중심점을 TSS의 위치로 예측한다.

2.2.3 *k*-mer에 대한 PWM

코어 프로모터 영역과는 달리 근접(proximal) 프로모터 영역은 공통적인 요소 및 구성을 찾아보기 힘들다. 이런 영역에 대해 *k*-mer의 통계적인 분포를 반영한 모델링 기법의 선례가 있으며[2,4,7,8,9], 본 논문에서는 DPF[2]에서 사용한 *5*-mer의 PWM을 이용한 프로파일링 기법을 적용하였다. *5*-mer의 PWM은 <그림 3>과 달리 4^5=1024의 행을 가지며,

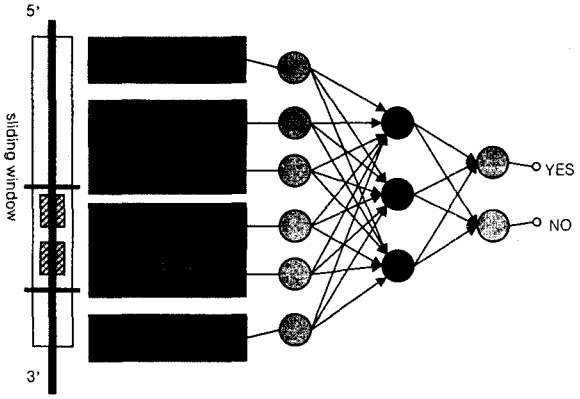

그림 4. PromSearch 구조개요. 250bp 크기의 영역을 분할하여 생성한 4개의 신호(signal)를 신경망을 이용해 통합하여 프로모터 여부를 결정한다.

길이가 L인 염기서열에 대해 다음과 같은 식을 통해 0과 1 사이의 값을 가지는 프로파일을 계산하게 된다.

$$S=\frac{\sum_{i=1}^{L-4} p_j{}^i \otimes f_{j,i}}{\sum_{i=1}^{L-4} m_j f_{j,i}}, \quad p_j{}^i \otimes f_{j,i}=\begin{cases} f_{j,i}, & \text{if } p_i = p_j{}^i, \\ 0, & \text{if } p_i \neq p_j{}^i \end{cases}$$

5-mer의 PWM은 근접 프로모터 150bp 영역 및 DPE (downstream promoter element)가 때때로 발견되는 35bp 영역의 프로파일 생성에 사용된다.

2.2.4 신경망

PromSearch는 임의 길이의 염기 서열에 대해, <그림 2>의 프로모터 모델을 바탕으로 <그림 4>와 같이 250bp 크기의 슬라이딩 윈도우를 이동시켜가며 윈도우의 4개 세그먼트에 대한 프로파일을 PWM을 이용해 생성한 후, 최종적으로 신경망을 이용해 프로모터 여부를 판별한다.

신경망은 기본적인 순방향(feed-forward) network, 즉 MLP(multi-layer perceptron)를 적용하며, Weka 패키지[10]를 이용해 학습하였다. 학습에는 PWM을 생성한 데이터와는 별도의 데이터를 사용한다.

3. 데이터 집합 및 성능 평가

3.1 PWM 생성

TATA-box와 Inr의 PWM은 EPD release 75의 척추동물 프로모터 집합을 입력으로 SSA의 PatOp 서비스를 실행하여 생성한다. 근접 프로모터 영역 및 DPE의 *5*-mer PWM 생성에는 EPD release 75의 인간 프로모터를 사용하며, 별도의 프로그램을 작성하였다. PWM 생성에 사용하는 FASTA 포맷의 서열에서 와일드 카드 'N'이 일정 비율 이상 포함된 경우에는 해당 서열은 사용하지 않았다.

3.2 신경망 학습

신경망의 학습에는 GENIE[11]에서 사용한 데이터 집합을 사용하였다. GENIE 데이터는 프로모터, 엑손, 인트론 세 종류의 서열의 집합이며, 프로모터 서열은 EPD release 50에서 추출한 565개 샘플, 엑손과 인트론 서열은 GenBank에서 추출한 890개, 4,445개의 샘플로 구성되어 있다.

이 데이터를 10-fold cross validation을 통해 학습하였으며, 히든 노드의 수는 3, 학습율 0.3, 모멘텀 0.2, epoch 500으로 설정하였다.

3.3 평가(evaluation) 데이터

Fickett & Hatzigeorgiou (1997) [12]의 리뷰 논문에서 사용한 데이터를 사용한다. 데이터는 24개의 TSS를 포함하는 전체 길이 33,120bp인 여러 서열의 집합으로 구성되어 있다. 이 논문에서는 다양한 방식의 프로모터 예측 프로그램간의 공정한 비교 방법을 제시하였으며, 당시 EPD에 포함되지 않은 새로운 인간 염기서열을 이용하여 여러 프로모터 예측 시스템의 평가를 수행하고 결과를 비교하였다. 본 논문에서는 이 리뷰 논문을 기준으로 PromSearch의 성능 평가 및 비교 결과를 제시한다.

4. 결과

성능 평가 척도는 민감도(sensitivity)와 특수도(specificity)를 기준으로 하였으며, 기존의 연구 결과와 비교를 위해 <표 2>에서는 [9]의 Table2를 참조하였다.

표준적인 10-fold cross-validation 및 Fickett의 데이터에 대한 평가 결과는 <표 1> 및 <표 2>와 같다. <표 2>에서 TP 및 FP 수의 결정은 Fickett의 기준[12]을 그대로 반영하였다.

GENIE 데이터에 대해서는 충분히 학습이 되었음을 <표 1>을 통해 확인할 수 있다. <표 2>를 보면 현재의 PromSearch 성능은 기존의 프로그램에 비해 아직은 미비하다. 높은 FP 문제는 해결하지 못했으나, 비교적 단순한 모델링을 통해 얻은 결과로서는 높은 coverage를 얻었다.

표 1. PromSearch의 sensitivity, specificity

평가방법	척도	결과
GENIE set (10-fold cross validation)	Sensitivity	67.0 %
	Specificity	91.2 %
평가 데이터 집합 (Fickett)	Sensitivity	43.2 %
	Specificity	88FP (1/376bp)

표 2. Fickett 데이터 집합에 대한 평가 결과 (PromSearch 이외의 방법에 대한 결과는 [9]를 참조하였다).

TP: true positive의 수, FP: false positive의 수

Coverage: 전체 24개의 TSS 중에서 각 프로그램이 정확히 찾아낸 비율

Method	TP	%TP of total matches	FP	TP/FP	Coverage (%)
Audic (1997)[13]	9	24	29	0.31	37
NNPP 2.1	14	19	59	0.24	58
TSSW (1997)	14	30	33	0.42	58
PromoterInspector (2001)	7	70	3	2.3	29
PromSearch	13	0.13	88	0.15	54

5. 결론 및 향후과제

PromSearch에서는 프로모터 영역을 4개의 부분으로 나누어 PWM을 이용한 프로파일을 생성하고, 이를 신경망으로 통합하여 프로모터를 예측하였다. 성능 향상을 위한 방안은 다음과 같다. 민감도를 높이기 위해서는, 좀 더 정확한 모델링이 필요하며, 이에 대해서는 현재의 버전에서는 고려하지 않았던 TATA/Inr PWM의 cut-off value를 반영한 신호 처리 과정을 추가하는 방법을 모색 중이다.

변별력을 높이고 특수도를 높이기 위한 방안으로는 non-promoter signal 추가 및 세분화, antisense strand를 포함한 프로파일링을 고려하고 있다. 그리고, 추가 프로모터 데이터(MGC, DBTSS, PRESTA, 기타)를 사용하고, 여러 데이터 집합을 이용한 평가가 병행되어야 한다.

감사의 글

본 연구는 ETRI의 '바이오 데이터 마이닝 및 통합관리 핵심 S/W 컴포넌트 개발' 과제에 의하여 지원되었음.

참고문헌

[1] U. Ohler, Computational promoter recognition in eukaryotic genomic DNA, PhD thesis, Technische Fakultät Erlangen-Nürnbeg, 2001.

[2] V. B. Bajic, A. Chong, S. H. Seah, and V. Brusic, An intelligent system for vertebrate promoter recognition. *IEEE Intelligent Systems*, July/August, 17 (4):64-70, 2002.

[3] M. G. Reese, Computational Prediction of Gene Structure and Regulation in the Genome of *D. melanogaster*. PhD thesis, University of Hohenheim, 2000.

[4] M. Q. Zhang, Identification of human gene core promoters in silico. *Genome Res.*, 8:319-326, 1998.

[5] 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.

[6] G. Ambrosini, V. Praz, V. Jagannathan, and P. Bucher, Signal search analysis server, *Nucleic Acids Research*, 31:3618-3620, 2003.

[7] G. B. Hutchinson, The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comp Appl Biosc.*, 12(5):391-398, 1996.

[8] V. Solovyev, and A. Salamov. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *In Proc Int Conf Intell Syst Mol Biol.*, 5:294-302, 1997.

[9] 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.

[10] Weka 3: Machine Learning Software in Java (http://www.cs.waikato.ac.nz/ml/weka/)

[11] D. Kulp, D. Haussler, M.G. Reese, and F.H. Eeckman, A generalized Hidden Markov Model for the recognition of human genes in DNA, *ISMB-96*, 1996.

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

[13] S. Audic and J. M. Claverie, Detection of eukaryotic promoters using Markov transition matrices. *Comput Chem.*, 21:223-227, 1997.