공통서열의 부분 정렬을 통한 전사인자 결합부위의 예측

윤주영[1O] 박근수[1] 임명은[2] 정명근[2] 박수준[2] 박선희[2] 심정섭[3]
[1]서울대학교 전기컴퓨터공학부, [2]한국전자통신연구원, [3]인하대학교 컴퓨터공학부
{jyyoon[O], kpark}@theory.snu.ac.kr, {melim, aobo, psj, shp}@etri.re.kr, jssim@inha.ac.kr

Prediction of transcription factor binding sites by local alignment of common sequences

Joo Young Yoon[1O] Kunsoo Park[1] Myung Eun Lim[2] Myung Geun Chung[2]
Soo-Jun Park[2] Sun Hee Park[2] Jeong Seop Sim[3]
[1]School of Computer Science & Engineering, Seoul National University
[2]Electronics and Telecommunications Research Institute
[3]School of Computer Science and Engineering, Inha University

요 약

유전자의 발현은 전사인자와 전사인자 결합부위의 결합에 의해 조절된다. 따라서 이러한 결합부위를 예측하는 것은 유전학 분야에서 중요한 이슈이다. 본 논문에서는 접미사 배열을 이용하여 전사인자가 결합할 것으로 예상되는 DNA 서열들의 공통서열을 추출하고, 이를 다시 입력 서열과 부분 정렬을 수행함으로써 전사인자가 결합하는 부위를 예측하는 알고리즘을 제시한다. 그리고 알려진 전사인자 결합부위를 가진 데이터로 실험한 결과를 통해 제시된 추출 방법의 성능에 대하여 논의한다.

1. 서 론

게놈 기술이 발전하면서 전사조절부위에 대한 연구가 활발해지고 있다. 유전자의 발현은 전사인자(transcription factor)와 DNA의 전사인자 결합부위(transcription factor binding site)의 결합에 의해 조절된다. 따라서 전사인자와 전사인자 결합부위에 대한 연구는 유전자의 기능을 예측하는데 있어 매우 중요한 연구 분야이다. 초기에는 생물학자의 실험에 의해 진행되었으나, 생화학적 실험에 의한 전사인자 결합부위의 예측은 많은 시간과 비용을 필요로 하기 때문에 이를 보완할 수 있는 *in sillico* 상에서의 계산적인 접근 방법이 점차 중요한 이슈로 떠오르게 되었다.

한편, Manber와 Myers에 의해 제안된 접미사 배열(suffix array)[1]은 텍스트 내에서 여러 패턴을 검색할 때 효과적인 인덱스 자료구조이다. 접미사 배열은 길이 n인 텍스트를 오직 n개의 값을 저장하는 배열로 전처리하고 빠른 시간 내에 패턴을 탐색할 수 있어, 대용량의 정보를 처리하는 생물 정보학 분야에서 다양한 문제에 응용되고 있다. 초기의 접미사 배열은 생성에 $O(n\log n)$ 시간이 필요하며 패턴 P를 탐색하는데 $O(|P|+\log n)$ 시간이 소요되었으나, 이후 활발한 연구가 진행되어 $O(n)$ 시간 생성 알고리즘[2,3,4]과 $O(|P|\log|\Sigma|)$ 시간 탐색 알고리즘[5]이 개발되었다.

본 논문에서는 이러한 접미사 배열을 사용하여 효과적으로 전사인자 결합부위를 예측하는 방법을 제시한다. 우선 입력 서열들을 접미사 배열로 전처리하고, LCP(longest common prefix) 정보를 이용하여 입력 서열들의 공통 서열을 추출한다. 다음으로 추출된 공통 서열들과 입력 서열들 사이에 부분 정렬(local alignment)[6,7]을 수행함으로써 각 공통 서열이 결합부위의 후보로 적합한지를 판단하고, 이를 통해 최종 후보와 그 위치를 얻는다.

이어지는 2장에서는 이러한 추출 방법과 관련된 기존의 연구에 대하여 소개한다. 3장에서는 우리가 제시하는 추출 방법에 대하여 설명한 후, 이에 대한 실험 결과를 보인다. 마지막으로 4장에서는 결론 및 앞으로 보완되어야 할 점을 논의한다.

2. 관련 연구

2.1 전사인자 결합부위 예측

전사조절부위에 대한 연구는 각각의 생물학자들이 실험실에서 실험 데이터를 정리하는 방법으로 시작되었다. 이후 발표된 결과들을 통합한 TRANSFAC[8]과 같은 데이터베이스가 구축되었고, 이를 기반으로 하여 전사조절부위를 예측하는 다양한 프로그램이 개발되었다. 그러나 이러한 기존의 데이터에 의한 접근 방법은 알려지지 않은 부위에 대한 예측이 불가능한 한계점을 가지고 있었다.

이를 보완하기 위하여 *in sillico* 상에서 새로운 접근 방법이 제안되었다. 알려진 데이터로부터 수리적, 통계적 모델을 만들어서 유사성을 검색하는 방법과 주어진 염기 서열만을 이용하여 전체적인 특성을 통해 예측하는 방법이다. 이러한 방법들을 사용하는 프로그램에는 CorePromoter[9], McPromoter[10] 등이 있다.

2.2 접미사 배열

주어진 텍스트에서 어떤 패턴을 찾는 방법은 크게 두 가지로 나뉜다. 첫 번째는 패턴을 전처리한 후에 텍스트에서 검색하는 것이고, 두 번째는 텍스트를 전처리하여 인덱스 자료구조를 만든 후에 패턴을 보면서 검색을 수행하는 것이다. 두 번째 방법은 일정한 텍스트에서 여러 개의 패턴을 연속적으로 검색할 때 효과적이다.

접미사 배열은 두 번째 방법에 해당되는 인덱스 자료구조이다. 기본적인 접미사 배열은 각각의 접미사들을 사전적 순서에 따라 정렬한 배열을 말한다. 그러나 다양한 문제를 효과적으로 해결하기 위해서, 추가적으로 인접한 두 접미사의 공통 접두사의 길이, 즉 LCP 정보를 저장한다.

2.3 부분 정렬

부분 정렬은 두 서열 사이의 유사한 부분의 쌍을 각각의 서열에서 찾는 문제이다. 많은 생물학적 문제에서 변이에 의해 서열이 부분적으로 달라지는 경우가 발생한다. 이러한 경우 패턴 검색 방법으로는 패턴이 서열에서 나타나는 위치를 찾을 수 없지만, 부분 정렬을 이용하면 찾을 수 있다.

대표적인 부분 정렬 방법으로는 Smith와 Waterman[6]과 Gotoh[7]에 의해 제안된 동적 프로그래밍을 이용한 알고리즘이 있다. 이 알고리즘에서는 서열 $A=a_1a_2...a_m$과 $B=b_1b_2...b_n$에 대하여, 다음과 같은 점화식을 이용하여 각 위치에서의 점수를 결정하고 가장 높은 점수를 가지는 위치를 부분 정렬의 결과로 얻는다.(일치 점수 α, 불일치 점수 δ, 초기 갭 페널티 γ, 갭 확장 페널티 μ)

$$H_{i,0}=H_{0,j}=0 \quad \text{for } 0\le i\le m, 0\le j\le n$$
$$H_{i,j}=\max\{0, H_{i-1,j-1}+s(a_i,b_j), C_{i,j}, R_{i,j}\}$$
$$\text{for } 1\le i\le m, 1\le j\le n$$

$C_{i,j}$와 $R_{i,j}$는 각각 A와 B에서 갭이 발생했을 때의 점수이다.

$$C_{0,j}=R_{i,0}=-\infty \quad \text{for } 0\le i\le m, 0\le j\le n$$
$$C_{i,j}=\max\{H_{i-1,j}-\gamma, C_{i-1,j}-\mu\}$$
$$\text{for } 1\le i\le m, 1\le j\le n$$
$$R_{i,j}=\max\{H_{i,j-1}-\gamma, R_{i,j-1}-\mu\}$$
$$\text{for } 1\le i\le m, 1\le j\le n$$

$s(a_i, b_j)$는 두 알파벳 a_i와 b_j 사이의 점수이다.

$$s(a_i,b_j)=\begin{cases}\alpha & \text{if } a_i=b_j\\ -\delta & \text{if } a_i\neq b_j\end{cases}$$

만일 부분적 정렬을 통해 일정 수준 이상의 유사도를 가지는 모든 위치를 얻고자 할 경우, 임계값을 정의하고 임계값 이상의 점수를 가지는 모든 위치를 결과로 얻는다.

3. 알고리즘 및 실험 결과

본 연구에서 제시하는 전사인자 결합부위 예측 알고리즘은 두 단계를 통해 수행된다. 먼저 주어진 서열들에서 공통서열을 추출하고, 이를 1차 후보라고 한다. 다음으로 1차 후보들을 각 입력 서열에 부분 정렬을 수행함으로써 후보와 충분히 유사한 서열을 가지는 입력 서열의 비율을 확인하고, 이를 통해 최종 후보를 결정한다.

3.1 1차 후보 추출

입력 서열들이 동일한 전사인자와 결합한다면 동일한 서열의 전사인자 결합부위를 가질 가능성이 높다. 따라서 입력 서열 중 일정 비율 이상의 서열에서 공통으로 나타나는 서열을 1차 후보로서 추출한다.

한 전사인자의 결합부위는 모든 유전자에서 동일한 서열을 가지지 않고, 조금씩 변형된 서열을 가질 수 있다. 따라서 공통 서열의 추출만으로는 유전자에 전사인자 결합 부위가 있는지를 정확하게 확인할 수 없다. 하지만 일정 비율 이상의 서열에서 공통으로 나타나는 서열은 전사인자 결합부위일 가능성을 가진다. 그러므로 공통 서열의 최소 길이 LEN, 전체 입력 서열 중 공통 서열이 나타나는 서열의 최소 비율 RTO_1의 두 인자를 사용하여 1차 후보를 생성한다.

효과적으로 공통 서열을 추출하기 위해 본 연구에서는 접미사 배열을 사용한다. 우선 각 서열의 끝에 서열의 알파벳에 존재하지 않는 종료 문자들($\#_1, \#_2, ...$)을 붙이고 모든 서열을 하나로 연결시킨다. 다음으로 연결된 서열의 접미사 배열과 LCP 정보를 생성한다. 생성된 접미사 배열과 LCP 정보를 이용하면 공통 서열을 추출할 수 있다.

만일 입력 서열에서의 전사 방향을 알지 못한다면, 일부 서열에서 전사인자 결합부위가 입력으로 받은 DNA 가닥(strand)이 아닌 다른 가닥에 존재할 수 있다. 따라서 이 경우에는 모든 입력 서열에 대하여 대응하는 반대쪽 서열(complement sequence)들을 포함해서 접미사 배열을 만들고, 입력 서열과 반대쪽 서열 중 하나 이상에서 나타나는 서열을 해당 입력 서열에서 나타나는 것으로 판단한다.

3.2 최종 후보 추출

이 단계에서는 전체 입력 서열 중 1차 후보가 나타나는 서열의 최소 비율 RTO_2를 인자로 하여 1차 후보로부터 최종 후보를 추출한다. 1차 후보가 최종 후보로 적합한지를 확인하기 위하여 각 1차 후보와 각 입력 서열에 대하여 부분 정렬을 수행한다. 1차 후보 P에 대한 부분 정렬의 임계값은 허용 초기 갭 개수 IG와 허용 확장 갭 개수 EG를 인자로 하여 다음 식과 같이 정의한다.

$$|P|\times\alpha-IG\times\gamma-EG\times\mu$$

만일 임계값 이상의 점수를 가지는 정렬 결과가 있다면 1차 후보가 이 입력 서열에서 나타나는 것으로 판단한다. 이를 통해 전사인자 결합부위의 최종 후보를 생성하며, 동시에 각 입력 서열에서 최종 후보가 나타나는 모든 위치를 얻을 수 있다.

3.3 실험 데이터

본 연구에서 제시하는 방법의 적합성을 평가하기 위한 실험 데이터로 Kato 등[11]이 제시한 모티프 조합(motif combination)을 사용하였다. Kato 등은 효모의 세포 주기 단계마다 전사인자 결합부위로 추정되는 모티프 조합과 이러한 모티프가 나타나는 유전자의 목록을 제시하였다. 이중에서 각각 두 개씩의 모티프가 나타나는 6개의 유전자 그룹의 서열을 입력으로 사용하여, 본 연구의 추출 방법을 다양한 인자로 실험하고 그 결과를 분석하였다. 각 그룹에서 나타나는 모티프와 그룹에 포함된 유전자의 수는 [표 1]과 같다. 각 유전자의 서열은 Wolfsberg 등의 웹 사이트[12]에서 600bp까지의 upstream서열을 얻어 사용하였다. 또한 이 서열에서 전사 방향은 알지 못하는 것으로 가정하였다.

[표 1] 각 데이터 그룹의 정보

그룹	1	2	3	4	5	6
모티프	ACGCGA CGCGAA	CGCGTC TGAAAC	CGCGAA TAAACAA	GTAAACA TTAGGAA	TGAAACA CCAGCA	AAACGC ATAATTA
유전자 수	75	37	10	14	8	10

3.4 실험 결과 및 분석

우선 LEN과 RTO_1을 변화시켜가며 실험을 수행하면서 발견된 최종 후보의 수와 [표 1]의 총 12개의 알려진 모티프에 대

한 탐색율을 조사하여 [표 2]와 [표 3]을 얻었다. [표 3]을 보면 LEN이 6 이하이면 RTO_1에 관계없이 알려진 모티프를 모두 찾아낸다. LEN이 7인 경우에는 길이가 6인 모티프를 대부분 찾지 못한다. 알려진 전사인자 결합부위의 모티프들이 대부분 길이가 6이상임을 감안하면 LEN의 값은 6이 적절하다고 판단된다. 이 실험에서는 모든 그룹에서 LEN이 6이하이면 RTO_1이 100%일 때에도 알려진 모티프를 모두 발견하고 있지만, 다른 실험 데이터를 사용할 경우에 존재할 수 있는 에러율을 감안하여 RTO_1은 80%~90%가 적당하다고 판단된다. 그룹별 분석을 위해 LEN이 6일 때의 그룹별 최종 후보의 수를 [표 4]에 제시하였다.

[표 2] 발견된 최종 후보의 수 (평균)

LEN \ RTO1	100	95	90	85	80
5	31.33	34.00	61.00	87.67	118.33
6	7.5	7.5	14.17	17.83	26.67
7	0.83	0.83	1.50	3.33	5.17

[표 3] 알려진 모티프에 대한 탐색율

(발견된 알려진 모티프의 수 / 전체 알려진 모티프의 수)

LEN \ RTO1	100	95	90	85	80
5	100	100	100	100	100
6	100	100	100	100	100
7	41.67	41.67	66.67	66.67	66.67

[표 4] LEN이 6일 때의 그룹별 최종 후보의 수

그룹 \ RTO1	100	95	90	85	80	합
1	2	2	6	7	16	33
2	2	2	3	6	12	25
3	10	10	21	21	40	102
4	7	7	20	26	26	86
5	15	15	15	27	27	99
6	9	9	20	20	39	97

다음으로 LEN=6, RTO_1=100%로 하여 부분 정렬의 임계값을 변화시켜가며 그룹마다 각각의 최종 후보가 서열에서 나타나는 평균 회수를 조사하였다. 이 실험에서 RTO_2는 결과에 영향을 주지 않았다. 부분 정렬의 인자는 $\alpha=4, \delta=1, \gamma=0, \mu=2$로 하였다. 이 결과는 [표 5]와 같다.

[표 5] 각각의 최종 후보가 서열에서 나타나는 평균 회수

(최종 후보가 나타나는 회수 / (최종 후보의 수 x 입력 서열의 수))

IG,EG \ 그룹	1	2	3	4	5	6	전체
1,1	7.32	8.70	25.70	22.59	23.01	24.67	18.21
1,0	3.43	3.12	8.44	7.94	8.29	8.90	6.59
0,0	1.30	1.41	2.09	2.12	2.43	2.89	2.00

최종 후보가 나타나는 평균 회수 역시 그룹 1, 2에서 비교적 적게 나타나는 모습을 보여주는데 이는 그룹 3~6에 포함된 최종 후보에 단순한 서열 구조(AAAAAA, AAAAAG 등)를 가진 최종 후보가 다수 포함되었기 때문이다. 이러한 최종 후보는 특정 위치에서 오버랩하여 여러 번 나타나기 때문에 최종 후보가 나타나는 평균 회수를 증가시킨다.

4. 결 론

본 논문에서는 접미사 배열을 이용하여 공통 서열을 추출하고 이에 대하여 부분 정렬을 수행함으로써 전사인자 결합부위의 후보 및 그 위치를 추출하는 방법을 제시하였다. 그리고 실험을 통해 수행에 적합한 인자를 제시하였다. 앞으로 단순한 서열을 가진 후보를 효과적으로 제거할 수 있는 방법을 도입한다면 보다 나은 결과를 얻을 수 있을 것으로 판단된다. 또한 각 최종 후보에 대한 부분 정렬 결과에 다중 정렬(multiple alignment)을 수행하여 최종 후보의 서열 구조를 분석하는 작업이 필요하다고 판단된다.

참고문헌

[1] U. Manber and G. Myers, Suffix arrays: A new method for on-line string searches, SIAM Journal on Computing, 22, 935-938, 1993.

[2] J. Kärkkäinen and P. Sanders, Simple Linear Work Suffix Array Construction, International Colloquium on Automata, Languages and Programming, LNCS 2719, 943-955, 2003

[3] D. Kim, J.S. Sim, H. Park, and K. Park, Constructing suffix arrays in linear time, Journal of Discrete Algorithms, Volume 3:126-142, 2005.

[4] P. Ko and S. Aluru, Space efficient linear time construction of suffix arrays, Journal of Discrete Algorithms, Volume 3:143-156, 2005.

[5] J.S. Sim, D. Kim, H. Park, and K. Park, Linear-time search in suffix arrays, Journal of KISS, 32(5), 255-259, 2005.

[6] T.F. Smith and M. S. Waterman, Identification of Common molecular subsequence, Journal of Molecular Biology, 147:195-197, 1981.

[7] O. Gotoh, An improved algorithm for matching biological sequences, Journal of Molecular Biology, 162:705-708, 1982.

[8] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, TRANSFAC: transcriptional regulation, from patterns to profiles, Nucleic Acids Research, 31(1), 374-378, 2003.

[9] M.Q. Zhang, Identification of human gene core promoters inSilico, 8(3), 319-326, 1998.

[10] U. Ohler, H. NieMann, G. Liao, and G.M. Robin, Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition, Bioinformatics, 17 Suppl 1, S199-206, 2001.

[11] M. Kato, N. Hata, N. Banerjee, B. Futcher, and M.Q. Zhang, Identifying combinatorial regulation of transcription factors and binding motifs, Genome Biology, 5:R56, 2004.

[12] T.G. Wolfsberg, A.E. Gabrielian, M.J. Campbell, R.J. Cho, J.L. Spouge, and D. Landsman,. Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae,. Genome Research 9:775-92, 1999.