공통서열 추출을 통한 전사인자 결합부위 예측

임명은° 심정섭 정명근 박선희
한국전자통신연구원
{melim°, simjs, cmk63697, shp}@etri.re.kr

Prediction of transcription factor binding sites by extracting common sequences

Myung Eun Lim° Jeong Seop Sim Myungguen Chung Sun Hee Park
Electronics and Telecommunications Research Institute

요 약

접미사 배열이나 접미사 트리는 대용량의 서열데이터를 효율적으로 검색, 저장할 수 있는 인덱스 자료구조로서 바이오인포매틱스와 같이 대용량 데이터의 처리, 분석이 필요한 분야에 이용될 수 있다. 최근 들어 접미사 배열에 대한 연구가 활발히 진행되어 접미사 배열의 효율적인 저장, 선형시간 생성 및 선형시간 탐색 알고리즘들이 개발되었다.

본 논문에서는 같은 전사인자가 결합할 것으로 예상되는 여러 개의 전사조절부위에 대한 DNA 서열들이 입력으로 주어졌을 때 전사인자가 결합하는 부위를 예측하는 방법을 제시한다. 이를 위해 최근에 제시된 선형시간 접미사 배열 생성 알고리즘을 이용하고 TRANSFAC과 EMBL 등의 DB를 이용하여 실험을 통해 본 논문에서 제시하는 방법의 정확도를 평가한다.

1. 서 론

접미사 트리[1]와 접미사 배열[2]은 문자열 알고리즘과 생물정보학 등의 분야에서 다양하게 응용되는 중요한 인덱스 자료구조이다. 접미사 배열은 쉽게 구현할 수 있는 장점이 있지만 접미사 트리에 비해 두 가지의 단점이 있다. 첫 번째 접미사 배열의 단점은 생성 시간이 오래 걸린다는 것이다. 길이 n인 텍스트 T에 대해 접미사 배열은 $O(n\log n)$의 생성시간이 필요하지만 접미사 트리는 상수 크기 또는 정수 알파벳에 대해 $O(n)$ 시간에 생성이 가능하다. 두 번째 접미사 배열의 단점은 탐색 시간이다. 접미사 배열에서 주어진 패턴 P의 탐색시간은 $O(|P| + \log n)$이고 접미사 트리의 패턴 탐색시간은 $O(|P|\log|\Sigma|)$로서 상수크기의 알파벳에 대해 접미사 트리가 더 빠르다.

최근 들어 접미사 배열에 대한 활발한 연구가 진행되어 위의 두 단점들이 극복되었다[3][4][5][6]. K□rkk□inen과 Sanders [3], Kim, Sim, Park, Park [4], Ko와 Aluru [5]가 각각 선형시간 접미사 배열 생성 알고리즘을 개발하였고, Sim, Kim, Park, Park [6]이 접미사 배열에서의 $O(|P|\log|\Sigma|)$ 시간 탐색 알고리즘을 개발하였다.

인간 게놈 프로젝트(human genome project) 수행 이후 유전체의 서열이 밝혀지면서 유전자 발현에 관여하는 전사조절인자 관련 연구에 관심이 커져 가고 있다. 전사조절(transcription regulation)에 대한 연구를 통해 '유전자 주석(gene annotation)' 즉, 유전자의 위치와 기능을 상세히 분석할 수 있고, '유전자 발현 조절(gene expression regulation)' 즉, 생체 조건에 따라 유전자의 발현 정도를 살펴봄으로써 유전자의 다양한 발현 가능성에 대한 연구를 진행할 수 있다. 전사조절부위 연구의 세부 분야로는 전사인자(transcription factor) 분야와 전사인자 결합부위(transcription factor binding site) 분야, 그리고 조절단백질분야 등이 있다. 전사인자 결합부위에 대한 연구는 이미 완료된 인간 염색체 지도와 대용량 실험인 DNA칩에서 얻어진 발현정보들과 더불어 유전자 기능예측을 위해 매우 중요한 연구 분야이다. 전사인자 결합부위는 80% 정도가 유전자의 5' upstream지역에 존재하며, 서열의 길이는 20bp 미만으로 알려져 있다[7].

전사인자 결합부위가 유전자에 비해 상대적으로 짧고 위치가 일정하지 않기 때문에 실험실에서의 전사인자 결합부위 예측은 시간과 비용이 많이 소요되어 매우 어렵다. 이를 보완하기 위해 *in silico* 상에서 다양한 접근이 진행되어 왔다[8][9]. 우선 알려진 일정분야의 데이터로 수리적, 통계적 모델을 만들어서 유사성을 검색하는 "search by signal"방법과 염기서열 클래스 전체적인 특성으로 예측하는 "search by content" 등이 이에 해당한다[9].

본 논문에서는 여러 개의 전사조절부위에서 같은 전사인자가 결합하는 부위를 예측하는 방법을 제시한다. 입력으로 주어지는 전사조절부위들은 서로 연관성이 높을 것으로 예상되는 서열들, 특히 같은 전사인자가 결합하는 부위를 포함할 것으로 예상되는 서열들이다. 예를 들어 DNA 칩에서 발현 패턴이 같은 유전자들의 upstream 부분이 입력 서열에 해당될 수 있다. 이러한 전사인자 결합부위의 예측을 위해 본 논문에서는 접미사 배열을 이용하여 전처리(preprocessing)하고 생성된 접미사 배열에 대한 LCP (longest common prefix) 정보를 이용하여 여러 서열에서 공통으로 존재하는 서열을 추출하는 방법을 사용한다.

먼저 2장에서 기존의 관련 연구에 대한 소개를 하고 3장에서 전사인자 결합부위를 예측하는 방법을 제시하고 이에 대한 실험 결과를 보인 후, 4장에서 결론을 맺는다.

2. 관련 연구

2.1 접미사 배열

주어진 텍스트에서 원하는 패턴을 찾는 정보 검색 방법은 크게 두 가지가 있다. 첫 번째는 패턴을 전처리하여 필요한 자료구

조를 만든 후 텍스트에서 검색을 수행하는 것으로 문서 편집기에서 원하는 단어를 찾는 경우를 예로 들 수 있다. 두 번째는 텍스트를 전처리하여 인덱스 자료구조를 만든 후 패턴을 보면서 검색을 수행하는 것으로, 책에서 색인을 보고 단어를 찾는 경우가 이에 해당한다. 두 번째 유형에 사용되는 대표적인 자료구조로는 접미사 트리와 접미사 배열이 있다.

Manber와 Myers [2]에 의해 제안된 접미사 배열은 각각의 접미사들을 사전적 순서에 따라 정렬시킨 후 그 순서를 배열 형태로 가지고 있는 자료구조이다. 접미사 배열은 접미사 트리에 비해 구조가 간단한 실용적인 모델이다.

접미사 배열에서 효율적인 패턴 탐색을 위해 LCP (longest common prefix) 정보가 이용되어 왔다. LCP는 접미사 배열에서 이웃하는 접미사들의 공통된 접두사의 길이를 나타낸다. 즉 LCP 배열 $L[i](1 \le i \le n-1)$는 i번째 접미사와 $i+1$번째 접미사의 공통 접두사의 길이를 나타낸다. 예를 들어, 주어진 텍스트가 $T = abbabaababbb\#$일 때, $A_T[3] = abaababbb\#$이고 $A_T[4] = ababbb\#$이다. ([그림 1] 참조) 이 때, $L[3] = 3$이다. 최근에는 LCP 정보를 이용하지 않고 접미사 배열에서 패턴을 탐색할 수 있는 알고리즘이 개발되었지만 LCP는 다양한 문제에 응용될 수 있기 때문에 중요한 정보이다. LCP는 선형시간에 생성이 가능하다[10]. 텍스트 $T = abbabaababbb\#$에 대한 접미사 배열 A_T와 LCP 배열 L은 [그림 1]과 같다.

i	$A_T[i]$	접미사	$L[i]$
1	13	$\#$	0
2	6	$aababbb\#$	1
3	4	$abaababbb\#$	3
4	7	$ababbb\#$	2
5	1	$abbabaababbb\#$	3
6	9	$abbb\#$	0
7	12	$b\#$	1
8	5	$baababbb\#$	2
9	3	$babaababbb\#$	3
10	8	$babbb\#$	1
11	11	$bb\#$	2
12	2	$bbabaababbb\#$	2
13	10	$bbb\#$	

[그림 1] 접미사 배열

2.2 전사인자 결합부위 예측

전사조절부위에 대한 초기의 연구는 생물학자가 실험실에서 개별적인 실험데이터를 정리하는 규모에서 시작되었다. 각각의 데이터는 문헌에 의해서 발표되었고, 이를 모으는 전문적인 큐레이터들의 노동집약적인 작업을 통해서 초기의 전사조절부위 연구가 진행되었다. 대표적인 예로 20여 년 전 러시아에서 시작한 TRANSFAC[11]을 들 수 있다. TRANSFAC은 전사인자, 전사인자의 결합부위, DNA-binding profiles의 데이터베이스로서, TRRD [12] 등의 공개된 자료와 논문을 기반으로 SITE, FACTOR 등의 6가지 정보를 체계적으로 구조화한 데이터베이스이다. 현재까지는 MatInspector [13] 등 대부분의 전사조절부위에 관한 프로그램들이 TRANSFAC의 데이터를 기반으로 개발되었다. 하지만 전통적인 큐레이션에 의한 접근은 데이터의 양이 매우 한정적이며 알려지지 않은 부위에 대한 예측이 어렵다.

이러한 초기 방식의 단점을 보완하기 위해 *in silico* 상에서 새로운 접근이 있었다. 모델링을 통한 검색방법과 유사한 종간의 유사성의 비교를 통한 방법이 있는데, Werner's group의 CoreInspector [14]와 MCPromoter [8]이 대표적이다. [14]는 "search by signal" 방법을 이용하며 [8]은 "search by content" 방법을 이용하는데 두 방식 중 "search by content" 방식이 조금 더 민감(sensitive)한 것으로 알려져 있다[15]. 유사한 종 사이의 유사성을 이용한 방식은 이론적인 기반이 완성되었으나 현재까지는 데이터가 부족한 상태이지만 각종 지놈 프로젝트가 완성됨에 따라서 점차 활성화되고 있다.

3. 전사인자 결합부위 예측 및 실험 결과

3.1 전처리 및 공통서열 추출

입력으로 주어지는 서열들은 서로 연관성이 높을 것으로 예상되는 서열들이다. 따라서 공통된 서열을 가질 가능성이 높을 것으로 가정한다. 이러한 공통 서열을 찾기 위해 본 연구에서는 접미사 배열을 이용한다. 각 서열들의 끝에 알파벳 ($\Sigma = \{A, C, T, G\}$)에 존재하지 않는 특수 문자들($\#_1, \#_2, \ldots$)을 붙이고 모든 서열들을 차례로 결합한 뒤 생성된 문자열에 대한 접미사 배열을 생성한다. 즉, 입력으로 주어지는 m개의 각 서열을 $S_1, S_2, \ldots, S_m$이라 할 때, $T = S_1\#_1 S_2\#_2 \cdots S_m\#_m$ ($\#_i, 1 \le i \le m$,는 Σ에 존재하는 모든 문자들보다 사전 순서가 작은 서로 다른 특수 문자이고, $i < j$에 대해 $\#_i < \#_j$)를 만들고 T에 대한 접미사 배열 A_T를 생성한다. 접미사 배열 생성 알고리즘은 최근 개발된 KÄrkkÄinen과 Sanders [3]의 알고리즘을 이용하였다. 이 알고리즘은 divide and conquer 기법을 이용하여 주어진 텍스트에 대한 접미사 배열을 선형시간에 생성하는 알고리즘이다. T에 대한 접미사 배열 A_T를 생성한 후 A_T에 대한 LCP 정보를 생성한다.

전처리 과정에서 생성한 접미사 배열과 이의 LCP 정보를 이용하여 공통서열을 추출한다. 전사인자가 결합하는 부위의 서열은 생물의 종, 유전자에 따라 서로 다를 수 있기 때문에 모든 입력 서열에서 충분한 길이의 공통서열이 발견될 것이라고 보장할 수는 없다. 따라서 입력 서열들 중 결합부위(공통서열)가 나타나는 서열의 개수의 비율을 나타내는 RTO, 그리고 결합부위의 길이 LEN에 대한 조건이 필요하다. 본 연구에서는 위의 RTO와 LEN 두 인자(parameter)를 변화시켜가면서 실험하여 각 인자에 따른 정확도를 비교한다.

3.2 실험 데이터

본 논문에서 제시하는 방법의 정확도 평가를 위해서는 같은 전사인자 정보와 이에 결합하는 결합부위들에 대한 관련 정보가 필요하다. 즉, 하나의 전사인자에 대해 이에 결합하는 (1)결합부위의 서열정보와 (2)결합부위를 포함하는 서열정보(upstream 서열정보)가 필요하다. (1)의 정보는 TRANSFAC Release 3.2 (1997-06-27)[11]의 "Factor" 정보를 이용하여 추출하였고, (2)의 정보는 TRNASFAC에 참조(reference)된 EMBL ID를 이용하여 EMBL Nucleotide Sequence Database (Release 75)[16]에서 300bp 길이를 얻어냈다. (1)과 (2)에서 얻어진 자료를 종합하여 만든 데이터는 사용한 데이터베이스들의 특성상 매우 많은 중복된 데이터들이 존재하게 된다. 이를 개선하기 위해 접미사 배열의 LCP와 EMBL의 유전자 정보를 참고하여 LCP가 비정상적으로 큰 데이터들은 중복된 자료로 가정하고

제거하였다. 이러한 방법으로 전사인자 10개에 대한 자료를 생성하고 이를 입력으로 사용하였다.

3.3 실험 결과

일반적으로 전사인자 결합부위의 길이가 20bp(base pair) 이내인 것을 감안하여 본 실험에서는 [표 1]과 같이 RTO와 LEN을 변화시켜가며 실험하였다. SAbuilder로 전사인자 결합부위 예측 결과를 검증하기 위해서, 예측 되어진 결과에서 실제로 올바른 예측 비율을 나타내는 Positive Probability Value (PPV)를 계산하여 [표 1]을 얻었다($PPV = \frac{TP}{TP + FP}$, TP: True positive, FP: False positive). 실제로 추정된 전사인자 결합부위의 길이 LEN이 4(bp) 이상이면서 RTO가 85% 이상인 경우에서도 30% 이상의 PPV 값을 나타냈다. 그러나 LEN을 4로 설정할 경우 false positive의 증가로 인해 실제 실험을 통한 검증이 매우 어려워진다. 프로그램을 수행하여 분석해 본 결과 LEN은 5-7bp가 적합하며, RTO는 65%~85%가 적절한 것으로 판단된다.

[표 1] Positive probability value

bp \ %	100	95	90	85	80	75	70	65	60	55
4	30.9	30.9	33.0	35.3	29.4	29.6	30.0	30.2	27.5	27.6
5	18.2	18.2	14.6	17.5	13.0	16.9	7.6	18.1	15.9	16.7
6	0	0	0	0	15.2	14.3	13.3	13.3	10.9	12.1
7	0	0	0	0	25.0	25.0	25.0	25.0	10.4	10.0
8	0	0	0	0	0	0	0	0	8.6	8.2
9	0	0	0	0	0	0	0	0	7.3	7.3

4. 실험 결과 및 결론

본 논문에서는 접미사 배열을 이용하여 공통 서열을 추출함으로써 전사인자 결합부위를 예측하는 방법을 제시하였다. 본 논문에서 제시되는 방법은 정확한 문자열 알고리즘(exact string matching algorithm)을 기반으로 하였다. 그러나 전사인자 결합부위는 같은 전사인자에 결합하더라도 종, 유전자에 따라 다양할 수 있으며 여러 부위가 서로 관련된 경우가 많으므로 향후 근사 문자열 알고리즘(approximate string matching algorithm)과 다중 서열 탐색(multiple pattern matching)의 적용이 필요하다고 판단된다.

참고문헌

[1] E.M. McCreight, A space-economical suffix tree construction algorithm, Journal of the ACM, 23, 262-272, 1976.
[2] U. Manber and G. Myers, Suffix arrays: A new method for on-line string searches, SIAM Journal on Computing, 22, 935-938, 1993.
[3] J. K□rkk□inen and P. Sanders, Simple linear work suffix array construction, International Colloquium on Automata, Languages and Programming, LNCS 2719, 943-955, 2003.
[4] D. Kim, J.S. Sim, H. Park, and K. Park, Linear-time construction of suffix arrays, Combinatorial Pattern Matching, LNCS 2676, 186-199, 2003.
[5] P. Ko and S. Aluru, Space efficient linear time construction of suffix arrays, Combinatorial Pattern Matching, LNCS 2676, 200-210, 2003.
[6] J.S. Sim, D. Kim, H. Park, and K. Park, Linear-time search in suffix arrays, Australasian Workshop on Combinatorial Algorithms, 139-146, 2003.
[7] J.W. Fickett and A.G. Hatzigeorgiou, Eukaryotic promoter recognition, Genome Research, 7, 861-878, 1997.
[8] M.Q. Zhang, Identification of human gene core promoters inSilico, 8(3), 319-326, 1998.
[9] M. Burset and R. Guigo, Evaluation of gene structure prediction programs, Genomics, 34, 353-367, 1996.
[10]T. Kasai, G. Lee, H. Arimura, S. Arikawa, K. Park, Linear-time longest-common-prefix computation in suffix arrays and its applications. Combinatorial Pattern Matching, LNCS 2089, 181-192.
[11] V. Matys, E. Fricke, R. Geffers, E. G□ßling, 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.
[12]E. Wingender, A.E. Kel, O.V. Kel, H. Karas, T. Heinemeyer, P. Dietze, R. Knuppel, A.G. Romaschenko, and N.A. Kolchanov, TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation, Nucleic Acids Research, 25(1), 265-268, 1997.
[13]K. Quandt, K. Frech, H. Karas, E. Wingender, T. Werner, MatInd and MatInspector - New fast and versatile tools for detection of consensus matches in nucleotide sequence data, Nucleic Acids Research, 23, 4878-4884, 1995.
[14]U. Ohler, H. Niemann, G. Liao, G.M. Rubin, Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition, Bioinformatics, 17 Suppl 1, S199-206, 2001.
[15]Gil-Mi Ryu, Mi-Ae Yoo, Development of drosophila promoter integration search system and visualization, Bioinformatics & Biocomplexity Research Center Pusan National University, 석사학위논문, 2002.
[16]G. Stoesser, W. Baker, A. Broek, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, F. Nardone, P. Stoehr, M.A. Tuli, K. Tzouvara, and R. Vaughan, The EMBL ncleotide sequence database: major new developments, Nucleic Acids Research, 31(1), 17-22, 2003.