유전자 알고리즘을 이용한 Promoter 예측

오민경, 김창훈*, 김기봉, 공은배, 김승목*
충남대학교 컴퓨터공학과 *생명공학연구소 유전체사업단

Promoter Prediction using Genetic Algorithm

Minkyung Oh, Changhoon Kim*, Kibong Kim, EunBae Kong, SeungMok Kim*
Dept. of Computer Engineering Chungnam National Univ. *KRIBB

요 약

Promoter는 transcript start site 앞부분에 위치하여 RNA polymerase가 높은 친화성을 보이며 바인딩하는 DNA상의 특별한 부위로서 여기서부터 DNA transcription이 시작된다. function이나 tissue-specific gene들의 그룹별로 그 promoter들의 특이한 패턴들의 조합을 발견함으로써 Specific한 transcription을 조절하는 것으로 알려져 있어 promoter로 인한 그 gene의 정보를 어느 정도 알 수가 있다. 사람의 housekeeping gene promoter들을 EPD(eukaryotic promoter database)와 EMBL nucleic acid sequence database로부터 수집하여 이것들 간에 의미 있게 나타나는 모든 패턴들을 optimization algorithm으로 알려진 genetic algorithm을 이용해서 찾아보았다.

1. 서 론

최근 들어 Human Genome Project의 시작으로 각 생물체의 방대한 유전정보들이 쏟아지고 있는 가운데 이들을 대량으로 통계적 분석을 해야할 필요성이 커지고 있다. 사람의 경우 3×10^9bp로 구성된 염색체를 가지고 있는데 그 중 2%만이 gene을 coding하고 있다. DNA 서열 중 이러한 gene들과 그의 signal 역할을 하는 promoter라는 specific한 부위를 효율적으로 밝혀내는 것이 중요하다.

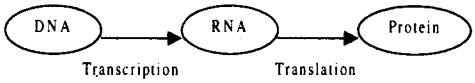

figure 1: Central Dogma

세포 내에 있는 염색체를 이루는 것이 DNA들로서 A, T, G, C의 네 가지의 조합으로서 그 생물체에 대한 모든 유전정보를 가지고 있다. 이중나선 DNA를 한 사슬에 해당하는 RNA로 합성하는 과정을 transcription이라 하고 또 이 RNA 중 messenger RNA(mRNA)상의 염기 서열을 틀로 하여 protein으로 합성하는 과정을 translation이라 한다. 이러한 기본적인 메커니즘을 central dogma라 한다.(fig. 1)

transcription이나 translation이 일어나게끔 조절을 해주는 여러 가지 RNA와 protein형태들이 존재한다. 그 중에서 transcription에서 중요하게 관여하는 것이 RNA polymerase이다. 이 RNA polymerase가 DNA 서열상에 높은 친화성을 보이며 binding하는 부분이 promoter이고 여기서부터 transcription이 일어나게 된다. prokaryote에는 한 종류의 RNA polymerase가 있고 eukaryote에는 세 종류의 RNA polymerase가 있다. 비교적 단순한 prokaryotic promoter에 대한 연구는 많은 성공적인 시도[1]가 있는 반면에 eukaryotic promoter는 자체의 복잡한 구조 때문에 그다지 많지가 않다[2][3]. prokaryote와 eukaryote의 차이는 간단히 핵의 유무에 있다. 세포 내에 핵이 존재하여 염색체가 핵 안에 보호되어 있는 고등한 생물체를 eukaryote라 하고 반대를 prokaryote라 한다.

eukaryotic promoter sequence에서 나타나는 transcription element(이하 TE) site(주로 6~8bp)들 중에 일반적으로 알려진 것 중 대표적인 것이 TATA box라든가 initiator가 있다. 그 외에 수백 가지의 알려진 TE들이 있는데 promoter sequence에는 그것들의 specific function이나 tissue에 따라 어떤 unique한 조합이 포함되어 있다고 알려져

있다. 이러한 TE들을 알아보기 위해 functionally specific promoter를 포함하는 sequence에서 의미 있게 나타나는 패턴들을, 이미 여러 multiple alignment[4][5]에 효과적으로 쓰여진 유전자 알고리즘을 이용하여 있는 대로 다 찾아보고 분석했다. 특히 사람의 housekeeping gene 20개에 대해서 실험을 했다. housekeeping gene은 발현했을 때 central dogma의 주요 기작에 관여하게 되는 gene들을 말한다.

2. 유전자 알고리즘(GA)

GA은 1975년 홀랜드의 저서 Adaptation in Natural and Artificial Systems에 처음으로 소개되었는데 natural genetics와 natural selection의 원리에 바탕을 둔 통계적 최적해 탐색 방법이다. 기존의 최적해 탐색이 국부 탐색이었는데 비하여 GA[6][7]는 여러 해를 동시에 탐색하는 전역 탐색을 함으로써 전역적인 최적해를 찾을 확률이 기존의 최적화 탐색에 비해 큰 것이 특징이다.

GA는 생물진화의 원리, 즉 선택도태나 돌연변이로부터 착안된 알고리즘으로서, 확률적 탐색이나 학습 및 최적화를 위한 한가지 기법이라고 간주할 수 있다. 첫 번째 단계에서 initial population을 생성해야 하는데 이 population의 멤버는 문제를 일차원 string으로 코딩을 한 개체들로서 각 string은 유전자형(genotype) 또는 염색체(chromosome)로서 참조되기도 한다. 보통 initial population은 랜덤하게 생성한다. 이렇게 생성된 initial population을 가지고 figure 2의 과정을 수렴에 이를 때까지 반복 수행한다.

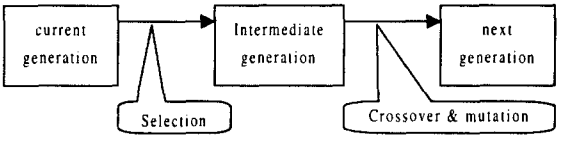

figure 3: generation process

GA는 기본적으로 세가지 종류의 유전자 조작, 즉 선택(selection), 교차(crossover), 그리고 돌연변이(mutation)를 사용한다. Selection을 적용하기에 앞서 개체가 다음 세대에 얼마나 살아남을 것인지를 측정하기 위한 수단으로 적합도 함수를 이용해서 그 개체에 대한 적합도를 구한다.

3. Promoter Prediction을 위한 알고리즘

Length가 L인 sequence들이 N개가 있을 때 window size, W의 길이로 자른 string들을 모은 data set을 만든다. 즉 이 data set에는 (L-W+1)×N개의 string들이 있게 된다. 그리고 GA의 operation을 적용할 initial population은 길이가 W인 string들을 data set으로부터 랜덤하게 P개를 추출한다.data set으로부터 추출하는 이유는 수렴속도를 줄이기 위함이다.

각 개체의 적합도(f_i)는 σ개의 substitution을 허용하여 data set으로부터의 frequency로 한다. 적합도가 구해지면 각 개체의 다음세대에 살아남을 수를 평균적합도(f)로 나눈 수로 정한다. 즉 f_i/f의 정수만큼 추가하고 소수부분만큼은 확률적으로 추가한다. 여기서 모여진 것이 figure 2의 intermediate generation에 해당한다. 이렇게 모여진 개체들을 두 개씩 조합을 하여 확률 Pc로 2점 crossover를 적용하고 각 개체의 base별로 확률 Pm으로 mutation을 적용하여 next generation을 만든다. 이러한 과정을 best적합도와 평균적합도의 비가 α에 이를 때까지 수행한다.

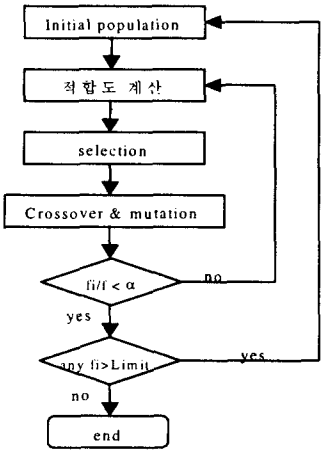

figure 3: GA process

수렴에 이른 population에서 유의한 적합도를 가지고 있는 개체를 선택한다. (L-W+1)×N개의 string들이 랜덤하다고 가정했을 때 σ개의 substitution을 허용하는 기대확률(P)을 이항분포를 이용하여 구한다.

$$P = \sum_{k=W-\sigma}^{W} \binom{W}{k} (0.25)^{W-k} (0.75)^{k}$$

그리고 기대되는 수 E, E=P×(L-W+1)×N이 된다.

정규분포로 transform하여 0.1%의 유의성으로 limit값을 다음과 같이 구한다[8].

$$Limit = 3\sqrt{(L-W+1)\cdot N \cdot P \cdot (1-P)} + E$$

이 Limit를 근거로 수렴에 이른 population에서 Limit보다 큰 적합도를 가진 개체들을 의미 있는 패턴으로 간주한다.

그리고 GA를 한번 수행시켜서 나온 패턴들 외에 적합도가 Limit보다 큰 패턴들이 빠졌을 가능성이 있기 때문에 GA를 수렴한 population에서 각 개체의 적합도가 Limit보다도 큰 값이 나오지 않을 때까지 계속해서 수행한다. 단 일단 패턴으로 나온 것들과 σ개의 substitution을 허용한 개체들의 적합도는 Limit보다 작은 값을 주어서 자연 도태되게 만든다. 그렇게 개체들의 적합도가 Limit보다 큰 것이 나오지 않을 때까지 GA를 반복 수행한다. 전체적인 스킴은 figure 3에서 보여진다.

4. 실험 및 결론

실험 대상의 housekeeping promoter는 Wataru가 [1]에서 실험을 했던 것과 같은 것을 EPD와 EMBL에서 추출하여 실험하였다. table 1에 나와 있는 것은 ID와 그것에 대한 키워드이다.

HSRNU1[V00591]	Human gene for small nuclear RNA U1.
HSU2A[K03023]	Human small nuclear RNA U2 gene.
HSU3PE[M14016]	Human U3 small nuclear RNA gene.
HSU4CA[M15957]	Human U4C small nuclear RNA gene.
HSHMG14A[M21339]	Human non-histone chromosomal protein HMG-14 gene.
HSHMG17G[X13546]	Human HMG-17 gene
HSNUCLEO[M60858]	Human nucleolin gene.
HSSNRNP3[M21253]	Human small nuclear snRNP E gene.
HSRPS14[M13934]	Human ribosomal protein S14 gene.
HSRIGA[M32405]	Human homologue of rat insulinoma gene.
HSACTBPR[Y00474]	Human beta-actin gene 5'-flanking region
HSHMGCOB[M15959]	Human HMG CoA reductase gene.
HSURODG[X06048]	Human URO-D gene for uroporphyrinogen decarboxylase
HSTPI[M10036]	Human triosephosphate isomerase mRNA.
HSPGK1[L00159]	Human phosphoglycerate kinase gene.
HSG6PD[X14520]	Human gene for glucose 6-phosphate dehydrogenase G6PD
HSSODIG[X01780]	Human superoxide dismutase gene.
HSYUBG[X04803]	Human ubiquitin gene
HSADPRF[M84327]	Human ADP-ribosylation factor 1 gene.
HSUBILP[03589]	Human ubiquitin-like protein gene.

table 1: List of housekeeping promoters

각 sequence는 promoter가 충분히 포함될 transcription start site로부터 200bp 앞쪽으로 하고 σ는 2로 하였으며 TE site가 주로 6~8bp임을 생각해서 W를 8로 두었다. crossover에 관한 확률 Pc의 값과 mutation에 관한 확률 Pm의 값은 수렴 접근 정도에 따라 Pc는 0.5와 0.2, Pm은 0.05와 0.01로 변화를 주었다. 뽑혀져 나온 string의 수는 대략 500개 정도이다. 많이 뽑혀 나오기는 하지만 이들의 대부분은 figure 4와 같이 같은 패턴의 양상을 보인다. 그러므로 이렇게 string들 간에 alignment를 해주고 그 결과적으로 나오는 것들을 선별하여 패턴으로 한다.

GGGCGGGG	GGGGGTGG
GGCGGGGG	GGGGTGGG
GGGGGCGG	GGGTGGGG
GGGGGGGC	GGGTGGCG
GGGGGGCG	CGGTGGGG
GGGGCGGG	GGTGGGGC

figure 4: GGGGGCGG, GGGTGGG와 유사한 결과를 보여줌

위의 두 가지 패턴(GGGGGCGG와 GGGTGGG) 외에도 사람의 housekeeping promoter의 패턴으로 밝혀진 것들 CCCCGCCC, GAGGCC, GATGGCGG, 등등이 일치하거나 거의 유사한 형태로 뽑혀져 나왔다.

유전자 알고리즘의 아주 기본적인 개념을 사용하여 특정한 function의 서열을 가지고 실험을 하였는데 이러한 패턴을 찾는 다른 논문의 방법들에 비해서 알고리즘이 아주 간단하면서도 신빙성 있는 결과를 보여주었다. 앞으로 specific function이나 tissue별로 그룹을 만들어 그룹간에 나타나는 차이를 보이는 연구가 필요하다.

Reference

[1] Tim Bailey and William E. Hart, "Learning Consensus Patterns in Unaligned DNA Sequences Using a Genetic Algorithm"

[2] Wataru Fujibuchi and Minoru Kanehisa, "Prediction of Gene Expression specificity by Promoter Sequence Patterns", DNA Research 4, 81-90 (1997)

[3] Dan S. Prestridge, "Predicting Pol II Promoter Sequences using Transcription Factor Binding Sites", J. Mol. Biol., 249, 923-932, 1995

[4] Ching Zhang and Andrew K.C.Wong, "A genetic algorithm for multiple molecular sequence alignment", CABIOS, vol. 13 no. 6 1997

[5] Cedric Notredame and Desmond G. Higgins, "SAGA: sequence alignment by genetic algorithm", Nucleic Acids Research, Vol. 24, No. 8, 1515-1524, 1996

[6] 기타노 히로아키, 유전자 알고리즘

[7] David Beasley, David R.Bull and Ralph R. Martin, An Overview of Genetic Algorithms, University Computing, 1993, 15(2) 58-69

[8] Martin Tompa, "An Exact Method for Finding Short Motifs in Sequences, with Application to the Ribosome Binding Site Problem", ISMB '99