SGT(Smoothed Good-Toulmin) estimator를 이용한 Dna sequeunce prediction

정호석
홍익대학교 회화과

요 약

본 연구의 주제는 Dna sequencing analyzing을 진행할 때 일부만 진행 된 검출 데이터를 기반으로 추가적인 분석이 진행되었을 때의 검출되는 Dna의 양을 예측하는 방식으로 SGT(Smoothed Good-Toulmin) estimator를 사용해보고 예측의 정확도를 분석하는 것이다. 분석은 각각의 염기서열이 검출될 확률이 랜덤으로 지정한 시뮬레이션을 기반으로 진행하였다.

Ⅰ. 서론

Dna 분석의 일련의 과정을 요약하자면 다음과 같다. 1. 다수를 Dna를 복제 시키고 유전자 가위로 잘라낸다. 2. 잘린 DNA의 가장 앞 부분에 특정 단백질을 부착한다. 3. 부착된 단백질은 자신이 결합한 Dna에 대해 전시적인 신호를 발생하고 그 뒤에는 다른 순서에 있는 Dna와 결합을 반복하여 단백질의 수명이 다하거나, 잘린 부위의 말단까지 읽어 더 이상 읽을 것이 없을 때까지 신호를 발신한다. 4. 단백질은 DNA를 이루는 4종류의 DDNTP(A,G,C,T)에 대해 각기 다른 전기적 신호를 발신하므로 발신하는 신호를 분석해 DNA 조각의 형질을 파악할 수 있다. 이렇게 1개의 단백질이 집어넣어 발신하는 일련의 신호를 리드라고 부른다. 5. 리드 단위로 DNA 조각들을 퍼즐을 맞추듯 위치를 매핑해서 분석하는 개체의 Dna sequence를 파악한다.

이 때문에 서로 다른 시간에 검출 된 두 리드는 항상 다른 것은 아니며 일부 공통적인 Dna를 가지거나 완전히 동일한 Dna들을 가지고 있을 수가 있다. 리드가 검출되는 확률은 단백질의 DNA와의 반응성이 다르기에 균등하지 않다. 이 때문에 서로 다른 두 종의 DNA에 대해 DNA의 차이 검사를 시행할 때 적은 수의 리드를 읽었을 때의 두 Dna에서 발견되는 차이 DNA의 비율과 많은 수의 리드를 읽었을 때의 DNA의 차이 비율은 상당히 다를 수 있다.[1] SGT는 Dna sequence 외의 다른 확률 모델에도 적용될 수 있다.

Good-Toulmin estimator는 무작위로 데이터 N개를 각각의 종이 수집 된 횟수에 대한 데이터가 주어진다면 M개의 데이터를 앞으로 더 수집했을 때 새로운 종이 몇 개가 더 발견될 것인가를 가늠할 수 있게 하는 수식이다. Good-Toulmin estimator는 수식의 특성상 M의 값이 N개 이하일 때에는 잘 예측하지만 N개를 초과하면 제대로 된 값을 예측 하지 못한다는 문제점을 갖고 있다. 선행된 연구들은 이 문제를 해결하기 위해 RFA, ETR, ZTNB의 방식을 적용해본 바가 있다.[2] 본 논문에서는 Smoothed Good-Toulmin의 방식으로 이 문제를 해결하려 하며 그 효과를 분석하고 있다.

Ⅱ.1장에서는 Good Toulmin이 도출된 배경과 수식이 갖고 있는 문제점에 대해서 Ⅱ.2장에서는 Good Toulmin의 문제점을 개선하기 위한 방법들과 Smoothed Good toulmin의 수식에 대한 설명을 다루고 있으며 Ⅲ장에서는 Smoothed Good toulmim의 방식을 적용한 시뮬레이션의 구현환경과 결

1) Timothy Daley and Andrew Smith (2013) Predicting the molecular complexity of sequencing libraries. Nat Methods. 2013 Apr;10(4):326
2) Timothy Daley and Andrew Smith (2013) Predicting the molecular complexity of sequencing libraries. Nat Methods. 2013 Apr;10(4):325-7. doi: 10.1038/nmeth.2375

과에 대해 고찰하고 마지막 Ⅳ장에서 본 연구의 결론을 도출하기로 한다.

Ⅱ. 배경이론

Ⅱ.1 Good-Toulmin estimator

Good-Toulmin의 수식이 발견된 유래는 다음과 같다. 1940년대에 영국 생물학자인 Corbet라는 사람이 2년간 섬에서 나비를 채집했다. 2년간 나비를 채집하여 나비 종류에 따라 잡힌 횟수를 기록했으며 영국으로 돌아간 Corbet은 R.A. Fisher라는 수학자에게 자신이 2년간 더 섬을 채류하려 하는데 똑같은 노력으로 나비를 채집하려들면 새롭게 발견될 나비가 몇 마리나 될 것인지 대해 예측해달라고 했다.[3)]

R.H.Fisher는 이 요청을 받고 예측할 수 있는 수학적 모델을 제시했고 후에 Good 과 Toulmin이 이 수학적 모델을 정리하여 발표한 수식을 Good-Toulmin estimator라고 부른다.

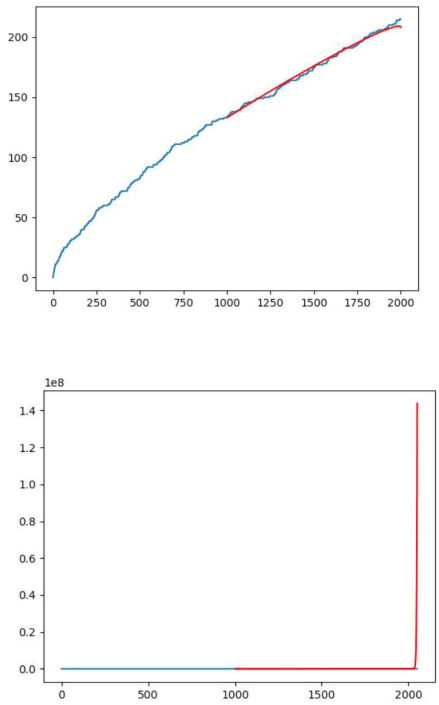
Good-Toulmin의 수식을 이해하기 위해 먼저 몇 가지의 선행적인 약속들을 살펴보도록 하자. 현재 상황은 분포가 알려지지 않은 집단에 대해 1개씩 임의 추출을 n번을 반복한 상황이며 앞으로 m번의 임의 추출을 더 진행할 것이라고 가정하자. t = m/n 으로 정의되며, $n_i(t)$는 m번의 시행을 진행하였을 때 j번 보여지는 표본의 갯수이다. 즉, $n_i(0)$란 현재의 상태에서 I번 발견된 종의 갯수를 의미한다.

$n_0(0)$ - $n_0(t)$란 m+n 번의 시행 끝에 한 번도 발견되지 않은 종의 수 - n번의 시행 끝에 한 번도 발견되지 않은 종의 수 = m번의 시행 동안에 새롭게 발견하게 될 새로운 종의 수를 의미하게된다. 구체적은 수식은 다음과 같다.

$$-\sum_{i=1}^{\infty}(-t)^i n_i(0)$$

이에 대한 증명은 Optimal prediction of the number of unseen species(written by Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu) p.13284을 참조하면 된다.

Good-Toulmin을 적용해보면 t의 값이 1을 초과할때 발산해버리는 성질을 확인할 수 있다.



동일한 확률 분포에 대해 N= 1000, M=1000(위),1050(아래) 일 때의 Good toulmin의 estimation 그래프

이는 Good-Toumin Estimator가 Power series($k(-t)^i$)이기에 t값이 1을 초과한 값을 가질 때 i값이 무한대가 되었을 때 발산해버리는 성질을 갖고 있기 때문이다.

3) Alon Orlitskya, Ananda Theertha Sureshb, and Yihong Wu(2016) Optimal prediction of the number of unseen species . PNAS, :13283

이에 대한 발산을 막는 전략으로 2가지를 제시할 수 있는데 첫 번째 방법은 Pade approximation 과 같이 주어진 수식을 근사시키되 발산 범위를 변경하는 방법이며 두 번째 방법으로는 Good-Toulmin에서 $(-t)^i$의 값은 i값이 짝수일 때에는 -∞로 발산하려는 성질을 i값이 홀수일 때에는 +∞로 발산하려는 성질을 갖고 있는데 이에 대한 가중치를 조절하여 t값이 1 이상일 때에도 발산하지 않고 잘 예측할 수 있도록 만드는 방법이 있다.

Ⅱ.2 SGT(Smoothed Good-Toulmin) estimator

SGT는 앞서 말한 두 가지 방법 중 두번째에 관련된 방식으로 기존에 있는 데이터가 가지는 중요도의 가중치에 변화를 주어 발산을 상쇄하겠다는 전략을 띄고 있다.

이에 대해 논문 Optimal prediction of the number of unseen species(written by Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu) pp.13285~13286 에 따르면 t가 1을 초과할시에 분포 파라미터가

$$r = \frac{1}{2t} log_e \frac{n(t+1)^2}{t-1}$$

Poisson(r) 분포를 따를 때에 효과적이라고 밝히고 있다.

Ⅲ. 시뮬레이션

Ⅲ.1 구현환경

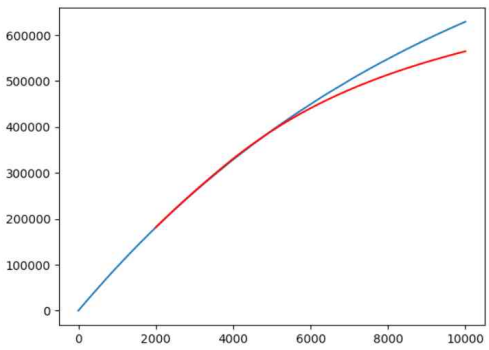
시뮬레이션은 Dna sequence prediction을 모델화한 상태에서 진행하였다. 분석하는 모델의 구현환경은 다음과 같다.

Dna의 갯수는 G(시뮬레이션은 10000000)개다. Dna를 분석에서 1개의 리드를 읽으면 L(100)개의 연달아있는 순서의 Base(Dna)를 알아낼 수 있다. 부위가 잘라지는 확률에 대해서는 랜덤이 적용됐다. N는 2000개의 리드를 읽고 모든 Dna는 오차가 없이 읽고 t는 4일 때의 시뮬레이션을 진행하였다.

이 프로그램의 x축은 read 단위로 되어있으며 y축은 x축 리드개의 검사에 따라 검출된 Dna의 갯수를 의미한다.

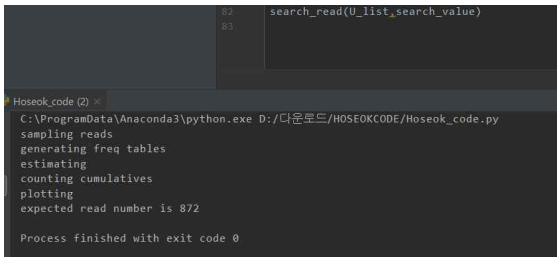
Ⅲ.2 시뮬레이션의 기능

Ⅲ.2.1 G,L,N,T의 값을 변경하고 리드 단위로 결과 그래프를 출력할 수 있다.



Ⅲ.2.2 G의 크기에 따라 4종류의 DDNTP가 임의로 배치하고 있다. 확률 분포에 따라 임의 추출하는 리드들의 DNA데이터와 위치정보가 txt 파일로 출력되어있으며, 그 txt파일을 입력데이터로 받아 SGT를 적용하여 예측하도록 코딩되었다.

Ⅲ.2.3 목표로하는 Dna 분석량을 설정하면 현재의 범위 내에서 몇 번째 추가적인 리드에서 찾아낼 수 있을 것인지 알 수 있도록 코딩되어있다.



Ⅲ.3 시뮬레이션 결과

t값에 따른 실제값과 시뮬레이션 값의 차이는 다음과 같았다.

	t=1	t=2	t=3	t=4
시뮬레이션 값(A)	329923	452601	553377	635443
SGT예측값(B)	330632.484	439098.416	510804.295	560642.743
A/B	0.998	1.031	1.083	1.133

Ⅳ. 결론

시뮬레이션을 통해 Dna sequence analyzing에 있어 SGT의 예측할 수 있음을 알 수 있었다. SGT 방식의 예측은 GT의 문제점이었던 T가 1이 이상인 지점에서도 발산하는 문제점을 보이지 않으나 T의 값이 커질수록 예측의 정확도가 떨어지고 있다.

논문 Predicting the molecular complexity of sequencing libraries(Written by Timothy Daley and Andrew Smith)을 보면 Pade approximation으로 접근하여 Dna sequence에 대한 Good-Toulmin의 발산 문제를 개선하였음을 알 수 있다. 본 연구는 동일한 목표에 다른 방향으로 접근을 한 것이므로 연구자로서는 두 가지의 방식 중에 무엇이 더 효과적인지에 대해 동일한 시뮬레이션으로 검사해보는 것도 의미 있는 비교가 될 것이다.

1.Timothy Daley and Andrew Smith (2013) Predicting the molecular complexity of sequencing libraries. Nat Methods. 2013 Apr;10(4):325-7. doi: 10.1038/nmeth.2375
2.Alon Orlitskya, Ananda Theertha Sureshb, and Yihong Wu(2016) Optimal prediction of the number of unseen species . PNAS, :13283-13288