올리고뉴클레오타이드 제작을 위해 효율적이고 차별적인 씨드를 고르는 방법에 대한 고찰

정원형○ 윤희근 박성배
경북대학교 컴퓨터공학과
whchung@sejong.knu.ac.kr, hkyoon@sejong.knu.ac.kr, sbpark@sejong.knu.ac.kr

A Study of Choosing Efficient Discriminative Seeds for Oligonucleotide Design

Won-Hyong Chung○ Hee-Geun Yoon Seong-Bae Park
Department of Computer Engineering, Kyungpook National University

1. 서 론

생물정보분야에서 올리고뉴클레오타이드(oligonucleotide)를 제작하는 문제는 시간을 많이 소모하는 문제이다. 이 문제를 해결하기 위하여 해쉬를 이용한 가속계산이 주로 쓰이고 있고 BLAST란 프로그램이 대표적으로 사용되고 있다. 생물정보학에서 지역정렬 문제는 DNA 또는 Amino acid 서열에서 정해진 임계값 이상의 유사도를 공유하는 영역을 찾는 문제이다. 반면 올리고 제작 문제는 DNA 또는 RNA 서열집합에서 개별 서열을 대표하는 짧은 길이의 서열조각을 찾는 문제이다. 두 문제는 공통적으로 답이 될 가능성이 희박한 서열부분을 제외시키는 사전작업을 통하여 계산속도를 향상시킬 수 있고, 가장 잘 알려진 방법으로 해싱이 있다. 이 방법은 시드의 형태와 길이에 따라 탐색범위가 달라지는 문제가 있다. PatternHunter에서는 불연속 매칭을 시드에 적용하여 이 문제를 개선하였다. 이후 시드를 개선하기 위한 연구가 지속적으로 이루어져 transition-constrainted seed, vector seed, BLAT seed 등의 시드들이 활용되고 있다. 이러한 시드들을 이용한 BLAST류의 프로그램들은 DNA서열의 특성에 따라 시드를 변형하여 해쉬를 개선하는 알고리즘을 적용하여 서열간의 유사도가 높은 부분을 찾는다. 그러나 이 프로그램들은 원래 올리고뉴클레오타이드 제작을 위해서가 아닌 지역정렬 문제를 해결하기 위한 방법들로써 발전하여 왔으므로 본 문제에 효율적인가에 대한 검증이 아직까지 이루어지지 않았다. 본 논문에서는 올리고 제작에서 효율적인 시드를 평가할 수 있는 잣대를 제시하고 이에 따라 잘 알려진 다섯 종류의 시드를 평가하여 어떤 시드가 올리고 디자인에 가장 적합한가를 제시한다.

2. 문제정의

올리고를 제작하는 문제에서 시드를 평가할 제약조건은 다음과 같이 지역정렬 문제와는 다른 이슈를 가진다. 첫째, 시드는 가능한 한 많은 올리고를 찾을 수 있어야 하고, 둘째, 시드는 올리고가 아닌 영역에서 잘못된 올리고 후보를 찾지 말아야 하며, 셋째, 올리고를 찾기 위해 사용되는 시드의 개수가 적을수록 좋다. 본 제약조건을 만족하는 잣대로서 '차별성'과 '효율성'을 제시하고 이를 동시에 고려하는 '효율적인 차별성'을 다음과 같이 정의한다.

효율적인 차별성을 정의하기 위해서 먼저 올리고 제작 시 시드를 이용하여 올리고가 매치하는 영역을 찾을 때 발생할 수 있는 경우들을 살펴본다. 그림 1에서는 목적하는 올리고 P0가 매치되는 영역 P1, P2를 시드 S0에 대한 탐색으로써 서열에서 찾고자 하는 경우를 추상적으로 표현하였다. S1이 P1에 겹친 경우는 목적하는 올리고를 찾은 경우이고, S2와 S3는 시드가 찾은 영역이 올리고와 매치하지 않는 영역인 경우이다. P2의 경우 올리고와 매치하는데 시드가 찾지 못한 경우이다.

차별성: 본 문제에서 차별성은 올리고가 매치하는 영역에서 시드를 발견하는 비율(precision)과 올리고가 아닌 영역에서 시드가 발견됨으로써 잘못된 올리고 후보를 찾는 비율(Recall)은 precision과 recall로 정의된다. 이를 통합하는 잣대는 precision과 recall의 조합평균 즉, F-measure로써 정의할 수 있다. 실제 올리고 제작 문제에서 precision과 recall에 대한 가중치가 다르게 부여될 수 있으므로 이를 조절하는 변수 α를 도입한 F_α를 차별성의 잣대로 사용한다.

효율성: 시드를 이용한 올리고 제작 시 두 가지 측면에서 시드 중복에 의한 효율성이 고려되어야 한다. 첫 번째는 해시를 구성하기 위하여 시드를 생성할 때 중복이 일어나는 비율($D = \frac{\text{number of generated seed hashes}}{\text{number of unique seed hashes}}$)이고, 두 번째는 올리고가 매치되는 영역에서 시드가 중복되는 비율($A = \frac{\text{number of seed hashes in oligos}}{\text{number of oligos}}$)이다. 효율성은 두 비율의 역함수에 각각 β와 γ의 가중치를 준 평균으로 정의할 수 있다. ($E_{\beta,\gamma} = \frac{\beta+\gamma}{\beta D + \gamma A}$)

효율적인 차별성: 앞서 정의한 효율성과 차별성은 서로 곱하여 하나의 잣대($G_{\alpha,\beta,\gamma} = F_\alpha \cdot E_{\beta,\gamma}$)로 사용가능하다.

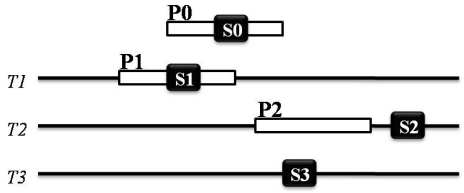

그림 1. 올리고 제작에서 시드를 활용할 때 발생가능 한 경우들: T1~T3은 서열, P0는 올리고, P1과 P2는 P0와 매치되는 T1과 T2에서의 올리고 위치들, S0는 P0에 속하는 시드이고 S1~S3은 서열에서 시드와 매치되는 위치들.

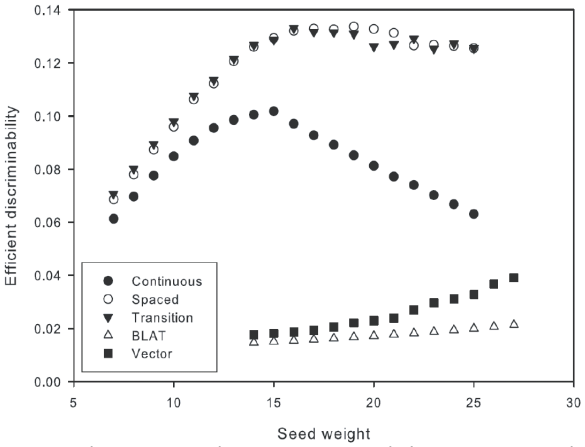

그림 2. 다섯 종류의 시드에 대한 효율적인 차별성 측정 결과

3. 실험 및 결과

본 논문의 실험을 수행한 프로세스는 다음과 같다. (1) 50길이의 DNA서열 1000개를 서열간의 유사도에 따라 고르게 준비한다. (2) 각 서열에서 모든 가능한 올리고를 제작하고 [6]에서 제시한 방법에 따라 평가하여 매칭위치를 기록한다. (3) 효율적인 차별성에 필요한 파라미터 α, β, γ 를 지정한다. 본 실험에서는 모두 1의 값을 주었다. (4) 다섯 종류의 시드들(continuous, spaced, transition-constrained, BLAT, vector)을 7에서 25까지의 무게에 따라 제작하여 실험에 사용한다. (5) 서열집합의 가능한 모든 위치에서 (4)에서 정한 각각의 시드에 대한 해시를 구성하고, 효율적인 차별성을 계산한다.

그림 2는 효율적인 차별성을 실험한 결과를 보여준다. 이 결과에서 무게 19의 spaced seed는 0.134의 값으로 가장 좋은 성능을 보였다. 전체적으로 transition-constrained seed와 spaced seed는 비슷한 패턴을 보였지만 무게 15이상의 시드에서 상대적으로 약간 낮은 성능을 보여주었다. continuous seed는 효율성이 고려되면서 spaced seed, transition-constrained seed보다 확연히 낮은 성능을 보였다. 그리고 무게 15까지 성능이 증가하다 이후 빠르게 감소함을 보였다. 이는 올리고의 매치를 결정할 때 길이 15이상 연속으로 매치가 되는 영역을 매치된다고 판단한 제약조건의 영향으로 무게 15까지는 전체적인 성능 향상이 있지만 이후 성능에 보탬이 되는 기술적 요인이 없기 때문이다. BLAT seed와 vector seed는 0.02 부근의 낮은 성능으로 시작하여 점차 성능이 증가함을 보였다. 이것은 두 시드들이 차별성과 효율성에서 낮은 성능으로 시작하기 때문이다. vector seed의 경우 무게가 증가함에 따라 성능이 상대적으로 많이 증가함을 보였다.

4. 결론 및 향후과제

본 논문에서 우리는 올리고 제작에 사용되는 시드의 성능을 측정할 수 있는 새로운 잣대를 제시하였다. 기존의 시드는 올리고 제작이 아닌 지역정렬에 특화되어 개발되었으므로 BLAT seed와 같이 지역정렬에서 좋은 성능을 보이는 시드가 의외로 올리고 제작에서는 효과적이지 않을 수 있다는 것을 보였다. 우리가 제시한 "효율적인 차별성"은 0과 1의 값으로 제한되고 값이 증가할수록 올리고 제작에 더 적합한 시드임을 보였다.

우리는 생물정보 분야에서 잘 알려진 시드 다섯 가지(continuous, spaced, transition-constrained, BLAT, vector)를 선정하여 차별성, 효율성, 그리고 효율적인 차별성에 대하여 측정하였고, 그 결과 효율적인 차별성에서 가장 좋은 결과를 보인 시드는 무게 19의 spaced seed, 차별성만을 고려했을 때 가장 좋은 결과를 보인 시드는 무게 12의 spaced seed이고 효율성은 시드의 길이가 증가함에 따라 계속 증가하는 결과를 얻었다.

본 논문에서는 기존에 알려진 시드에 대해서 평가하는 작업을 수행하였지만 차후에는 효율적인 차별성을 바탕으로 올리고 제작에 가장 적합한 시드를 새롭게 제작하여 제시하는 작업이 필요하다. 그리고 시드의 개선으로 실제 생물학 실험에서 얼마나 더 좋은 결과를 얻었는가를 측정하는 작업이 필요하다. 또한 입력 서열에 따라 시드를 변형하여 기존의 올리고 제작 프로그램의 성능을 높이는 툴의 제작 또한 가능하다.

감사의 말

본 논문은 교육과학기술부의 재원으로 BK21-IT 프로그램의 지원을 받아 수행된 연구임.
본 논문은 지식경제부 및 정보통신진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행 되었습니다. (A1100-0601-0102)