Fragment Assembly를 위한 시스템의 설계 및 구현

김명선[1] 정철희[1,2] 박현석[1,3]
㈜ 마크로젠[1]
고 려 대 학 교 컴 퓨 터 학 과[2]
세 종 대 학 교 컴 퓨 터 공 학 과[3]
{keystore}@macrogen.com
cholhee@formal.korea.ac.kr
hspark@cs.sejong.ac.kr

Design and Implementation of the genome-level fragment assembly system, Mater

Myung-Sun Kim[1] Chol-Hee Jung[1,2] Hyun-Seok Park[1,3]
Macrogen[1]
Dept. of Computer Science & Engineering , Korea University[2]
Dept. of Computer Science & Engineering, Sejong University[3]

요 약

지금까지 인간이나 다른 생물체의 전체 유전체 염기서열을 밝혀내는 작업은 크게 세가지 방법으로 진행되었다. Clone-by-clone approach, sequence tagged connector approach, random shotgun approach[1]가 그것인데 마지막의 random shotgun approach는 fragment assembly problem을 비롯한 여러 가지 전산학적인 문제들을 수반한다. 미생물체의 전체 염기서열을 random shotgun approach를 이용하여 밝혀낼 때 몇 가지 전산학적인 문제가 테크닉이 필요하며 그 중에서도 서열간의 forward, reverse의 mating 정보를 이용하는 것이 매우 중요하다. 본 논문은 이러한 mating 작업을 한 눈에 볼 수 있게 하는 소프트웨어 패키지 "Mater"에 대해 소개하고자 하며 그 의미에 대해 논하고자 한다.

1. 서론

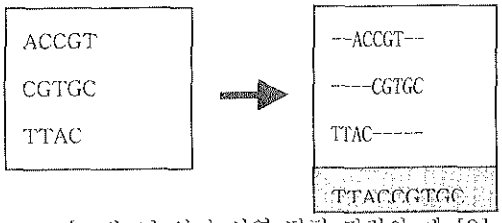
Whole-Genome random shotgun 방법은 어떤 한 생물체의 유전자 염기 서열을 밝혀내는 genome-project에서 오늘날 매우 많이 사용되는 방법이다. 이는 기존의 다른 방법들보다 훨씬 빠른 시간 안에 매우 큰 유전체의 염기서열을 분석할 수 있는 방법이다. 그러나 여기에는 전산학적, 통계학적인 방법들이 그 바탕에 깔려있다. 전산학적인 뒷받침 중 대표적인 것이 fragment assembly problem인데 이는 NP-Hard problem중 하나인 SCS-problem (Shortest Common Super-string problem)이다[2]. 이를 위해선 또 다른 전산학적 방법을 사용하는데 오차를 허용하는 pattern matching이 그것이다. 이러한 전산학적 문제에 대한 해결책은 '생물정보학(Bioinformatics)'이라는 새로운 학문 분야를 탄생시켜 활발히 연구중이다. 본 논문에서는 random shotgun 방법에 대해 간략히 설명 한 후, fragment assembly 작업을 가능하게 해 주기 위해 필요한 여러 소프트웨어 툴 중 저자가 Java로 개발한 Mater 소프트웨어 패키지에 대해 설명하고자 한다.

2. The random shotgun approach

Random shotgun 방식은 Frederick Sanger를 중심으로 고안된 길이가 긴 유전체의 염기서열을 밝혀내기 위한 한 방법이다. 이는 긴 유전체 염기서열을 임의의 위치에서 여러 개의 짧은 조각들로 나눈 후 그 조각들 사이에 중복되는 부분들을 이용하여 다시 원래의 긴 서열을 만들어내는 방법이다.

이런 방법으로 얻은 염기 서열 조각들은 다중 염기 서열 정렬 과정을 통해 몇 개의 긴 염기서열(contig)로 합쳐진다. 이 과정에서 두 염기 서열 단편간의 유사도(homology)는 PAM이나 BLOSUM같은 Scoring-matrix를 통해 계산된다.

[그림 1]의 왼쪽은 각 염기 서열의 단편, 오른쪽 위는 다중 정렬(multiple alignment) 된 모습이고 오른쪽 아래는 다중 정렬을 통해 얻은 하나의 긴 염기서열(contig)이다.

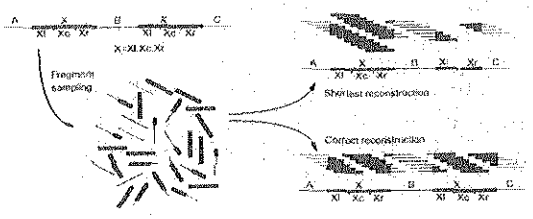


[그림 1] 염기 서열 단편 정렬의 예 [3]

3. Random shotgun 방식에서 해결해야 할 어려운 점들

1) 반복적인 염기서열

Random shotgun 방식은 유전체의 긴 염기서열을 매우 많은 수의 짧은 조각들로 나눈 후 그것들을 다시 하나의 긴 염기서열로 합친다. 그러나 생물체의 염기서열 중에는 일정 길이의 염기서열이 반복적으로 나오는 곳이 있다(repeated regions). 이런 경우엔 다음과 같이 염기 서열 조각들이 비정상적으로 결합하는 경우가 발생할 수 있다.



[그림 2] repeated region[1]

단순히 가장 짧은 superstring을 생성하는 경우 잘못된 contig를 만들게 된다.

2) 불충분한 염기 서열 조각(Lack of coverage)

6-fold가 넘는 충분한 수의 염기 서열 단편들을 분석했더라도 여전히 서열 단편들이 만들어지지 않은 부분이 남아있게 된다(gap). 이러한 부분은 생물학적으로 cloning-vector로 단편들을 복제할 수 없는 부분(un-clonable sequence)이거나 몇 가지 기계적인 이유에 의해 해당 부분의 염기 서열 조각이 생성되지 않은 부분이다.

3) 서열 분석 에러

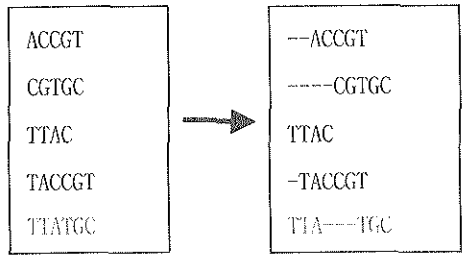
전자동 염기 서열 분석기를 통해 분석될 결과에도 에러가 있을 수 있다. 정교하게 조절된 실험 환경에서 처음 500bp정도 까지에서는 에러율이 약 1%미만의 확률로 일어나지만 그 이후에는 급속도로 증가 하여 650bp 정도에서부터는 에러율이 15%가 넘어 가기도 한다. 각 염기 서열 조각들의 끝 부분끼리의 유사성을 통해 하나의 긴 킨티그를 만들어 가는 염기서열 조각 재결합 문제(fragment assembly problem)에서 각 조각들의 끝부분에 에러가 많다면 잘못된 컨티그를 생성할 가능성이 매우 커지게 된다.

4) 방향성

DNA염기서열은 서로 상보적인 두 가닥의 서열이 나선형 구조를 이루고 있어서 분석한 염기 서열 조각의 실제 방향이 어떤 것인지 알아내기가 매우 어렵다. 이런 이유로 여러 개의 염기 서열 조각을 하나의 긴 컨티그(contig)로 재결합할 때 각각의 서열 조각들끼리의 유사성 검사를 최소 두 번씩 해야 하는데 이는 문제의 시간 복잡도를 증가시키게 된다.

5) 키메라(chimera)

마지막으로 서로 멀리 떨어진 곳에 있는 두 개 이상의 짧은 서열 조각이 서로 결합하여 마치 하나의 서열 단편처럼 보이게 되는 경우도 생긴다.(chimeric fragment).

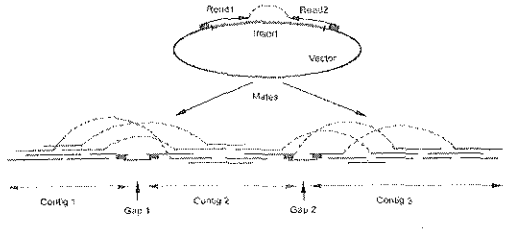


[그림 3] 비정상적 염기 서열 단편 (chimeric fragment)[4]

가장 아랫쪽에 있는 염기 서열 조각의 경우 서로 멀리 떨어진 두 염기 단편의 일부분이 하나로 합쳐진 것임을 알 수 있다. 이러한 잘못된 서열 조각들에 대한 고려 없이 일반적인 fragment assembly problem에 대한 알고리즘을 적용한다면 정체 불명의 생물체가 만들어질 것이다.

4. Mating 작업

위에서 말한 Whole-Genome random shotgun approach[5]에서 선결해야 할 몇 가지 문제에 대한 해결 방안으로 각 염기 서열조각을 다른 것과 짝을 이루도록 하는 것이 있다.



[그림 4] mate[1]

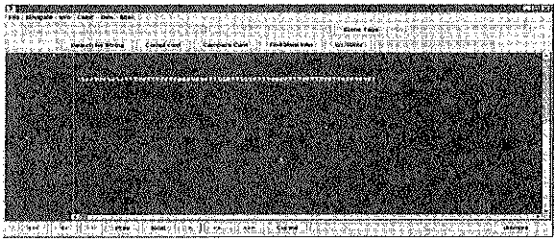
cloning vector에 삽입된 유전자 조각(clone)의 염기 서열을 양쪽에서부터 읽어나가면 동일한 유전자 조각에 대한 염기서열 분석 결과물이 두 개씩 짝을 이루게 된다. (그림 4)에서와 같이 각 염기서열 조각들이 두 개씩 짝을

이루게 되면 그것을 통해 많은 정보를 추가적으로 알 수 있게 된다.
먼저 짝을 이루는 두 개의 염기 서열 조각 사이의 대략적인 거리를 알 수 있다. 각 유전자 조각들은 길이가 2k나 10k가 되도록 하였으므로, 짝을 이루는 두 염기서열 조각의 길이를 각각 *l1,l2*라 했을 때 그 둘 사이의 거리는 약 2000-(*l1*+*l2*) 혹은 10000-(*l1*+*l2*) 가 된다. 직관적으로 이러한 거리 정보는 일정 염기 서열이 반복적으로 나타나는 부분(repeated region)을 재구성할 때 이용될 수 있다. 즉, 만약 여러 개의 염기 서열 조각으로 하나의 긴 컨티그를 만들었을 때, 짝을 이루는 두 개의 염기 서열 조각 사이의 거리가 계산치보다 더 짧다면 이는 잘못 구성된 컨티그라 할 수 있다. 이런 경우 두 염기 서열 조각 사이의 거리를 계산치 만큼 늘여 놓으면 반복적으로 나타나는 염기 서열 부분이 원래의 그것과 같이 정상적으로 만들어질 수 있다. 컨티그들 사이의 순서 정보도 짝을 이루는 두 염기 서열 조각을 통해 얻을 수 있다. Random shotgun 방법으로 유전체에 대한 염기서열 정보를 얻으려 할 때 부딪히는 어려운 점인 불충분한 염기 서열 데이터(Lack of coverage)로 인해 6-fold이상의 충분한 염기 서열 조각을 분석해도 일부분은 여전히 분석되지 않고 남아있게 된다. 이러한 갭(gap)부분은 별도의 방법을 통해서 메꿔야 하는데, 이 때 각 컨티그들의 순서 정보가 매우 중요하게 쓰인다. (그림 4)에서 볼 수 있듯이 컨티그들의 순서는 짝을 이루는 두 염기 서열 조각을 통해 얻어낼 수 있다.

5. Mater 소프트웨어 패키지의 설계 및 구현

1)컨티그파일 열기 (Open Contig)

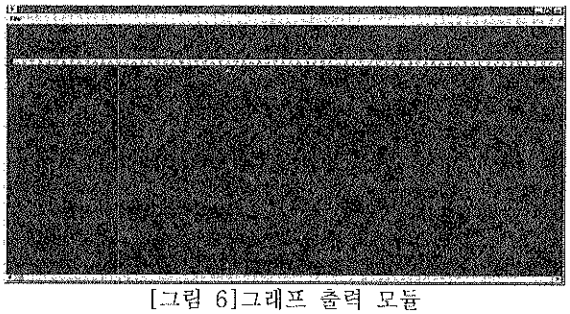
염기서열의 단편들을 다중정렬하고 이를 통해 얻은 하나의 긴 염기서열(contig)을 화면에 출력한다. 화면에 나타난 ATGC는 염기를 구성하는 아데닌(A, Adenine), 티민(T, Thymine), 구아닌(G, Guanine) 그리고 시토신(C, Cytosine)을 의미한다.



[그림 5] 컨티그 파일 출력 모듈

2)그래프 그리기

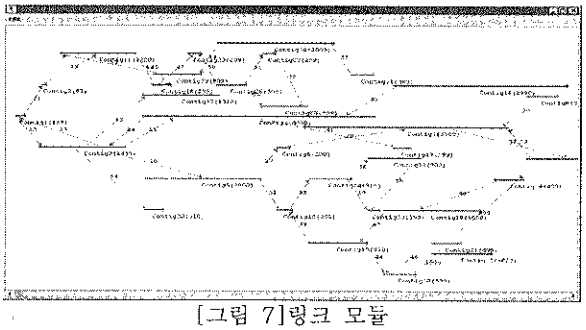
Open Contig 메뉴가 선택되어진 상태에서 화면에 나타난 리드(각 문자열) 중 하나를 선택한 후 특정염기(한 문자)에서 오른쪽 마우스를 클릭했을 때 보여지는 화면이다.



[그림 6]그래프 출력 모듈

3)링크 (Link)

긴 유전체 염기서열을 임의의 위치에서 여러 개의 짧은 조각들로 나눈 후 그 조각들 사이에 중복되는 부분들을 이용하여 다시 원래의 긴 서열을 만들어내는데 링크는 각 염기서열조각을 다른 것과 짝을 이루도록 하는 것이다.



[그림 7]링크 모듈

6. 결론

1998년 미국의 Celera Genomics란 한 벤처회사는 Whole-Genome random shotgun 방식으로 2001년까지 인간 유전체의 전체 염기서열을 밝히는 것을 목표로 하여 또 하나의 HGP를 시작하였다. 본 논문에서는 Random shotgun 방식에서 매우 중요한 Mate 정보에 대한 GUI, 즉 "Mater"를 제작함으로써 이 분야의 연구에 유용한 툴을 제작하는데 기여했다고 생각한다.

7. 참고 문헌

[1]. Gene Myers, "Whole-Genome DNA Sequencing"
[2]. D. Gusfield, Algorithms on strings, trees, and sequences - Computer Science and Computational Biology, 1997.
[3]. J. Meidanis / J. C. Setubal, "Introduction to Computational Molecular Biology", pp.107-108.
[4]. J. Meidanis / J. C. Setubal, "Introduction to Computational Molecular Biology", pp.108-109.
[5]. R.D. Fleischmann *et al.*, "Whole-Genome Random Sequencing and Assembly of *H.Influenzae*," Science, Vol. 269, No. 5,223, 1995, pp.496-512.