Random shotgun 방법을 이용한 생물체의 염기서열 분석

정철희 윤경오 박현석 최진영
고려대학교 컴퓨터학과
세종대학교 컴퓨터공학과
(주)마크로젠
{chjung, choi}@formal.korea.ac.kr
{hspark, yoonko}@macrogen.com

Whole-Genome Sequencing by the random shotgun approach

Chol-Hee Jung[1,3] Kyong-Oh Yoon[3] Hyun-Seok Park[2,3] Jin-Young Choi[1]
Dept. of Computer Science & Engineering, Korea University[1]
Dept. of Computer Science & Engineering, Sejong University[2]
Macrogen, Corp.[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. 이 논문은 저자들의 국내 최초로 미생물체의 전체 염기서열을 random shotgun approach를 이용하여 밝혀낸 경험을 바탕으로 그에 따르는 문제인 fragment assembly problem에 대해 소개하고 그에 수반되는 몇 가지 전산학적인 문제와 몇 가지 해결책에 대해 설명하려 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[6]. 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 논문의 저자들 또한 총 염기 서열의 길이가 약 2.4Mbp인 한 미생물 유전체의 전체 염기 서열 중 98,9%를 Random shotgun방법을 적용하여 국내 최초로 분석하였다. 이 논문은 이러한 know-how를 바탕으로, random shotgun 방법에 대해 간략히 설명한 후, 한 미생물체 유전체의 98,9%를 밝혀내는데 사용한 fragment assembly 작업에 따른 몇몇가지 난제들, 그리고 그 해결책에 대해서 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.

Random shotgun 방식으로 전체 염기 서열을 찾는 과정은 다음과 같다.

1) 초음파등을 이용하여 전체 염기서열을 여러개의 짧은 조각들로 나눈다.

2) 너무 길거나 너무 짧은 조각들을 제거하여 각 조각들의 편차가 10% 이내가 되도록 한다. 일반적으로 2kbp나 10kbp정도 되는 염기 서열 조각을 선별하는데, 2kbp와 10kbp의 비율이 약 4:1 정도 되도록 한다.

3) 2)번 과정을 거쳐 선별된 조각들을 vector-cloning 과정을 통해 똑같은 단편들을 대량 복제(cloning)한다.

4) 3)번까지의 과정을 거쳐 얻은 수많은 클론들의 염기서열을 전자동 염기서열 분석기를 통해 분석한다.

위 과정을 여러 번 반복하여 서열 조각들의 길이의 합이 전체 염기 서열을 충분히 포함하도록 한다.

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.

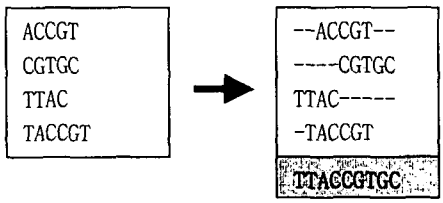

그림 1. 염기 서열 단편 정렬의 예 [2].
왼쪽이 각 염기 서열의 단편, 오른쪽 위가 다중 정렬(multiple alignment)된 모습, 오른쪽 아래가 다중 정렬을 통해 얻은 하나의 긴 염기서열(contig)이다.

현재 전자동 염기서열 분석기를 통해 한번에 분석할 수 있는 염기서열의 길이는 평균적으로 약 500bps(base pairs)다. 그리고, 통계학적으로 볼 때 한 가닥의 긴 염기서열을 임의의 위치에서 조각들로 나눴을 때, 조각들 수에 따라 전체 조각들이 분석하고자 하는 전체 염기서열을 포함하는 정도는 포아송 분포를 그린다[3]. 따라서 전체 염기 서열 수 중에서 분석되지 않은 염기서열 수는 다음과 같이 나타낼 수 있다(그림2).

$$P_0 = e^{(-LN/G)}$$

P_0 : 분석되지 않은 염기 수의 비율.
L : 클론들의 평균 길이.
G : 전체 염기서열의 길이.
N : 클론 수.

그림 2. 분석되지 않은 염기 수.
1-fold는 각 염기 서열 조각들의 전체 길이를 분석하고자 하는 전체 염기서열의 길이로 나눈 값이 1인 경우이다.

위 공식에 의하면 각 염기 서열 단편의 길이가 약 500bp로 일정하다고 할 때 6-fold만큼의 염기 서열 조각을 분석하면 $e^{(-6)}$는 약 0.002478전체 길이의 99.99%이상에 대한 염기서열을 얻을 수 있게 된다[5].

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 염기서열 조각들이 비 정상적으로 결합하는 경우가 발생할 수 있다(그림 3).

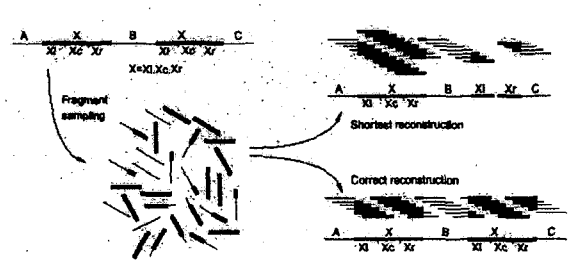

그림 3. repeated region[1]
단순히 가장 짧은 superstring을 생성하는 경우 잘못된 contig를 만들게 된다.

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(unclonable 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 증가하여 650bp 정도에서부터는 에러율이 15%가 넘어가기도 한다. 각 염기 서열 조각들의 끝 부분끼리의 유사성을 통해 하나의 긴 콘티그를 만들어 가는 염기 서열 조각 재결합 문제(fragment assembly problem)에서 각 조각들의 끝부분에 에러가 많다면 잘못된 콘티그를 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 콘티그

(contig)로 재결합할 때 각각의 서열 조각들끼리의 유사성 검사를 최소 두 번씩 해야하는데 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).

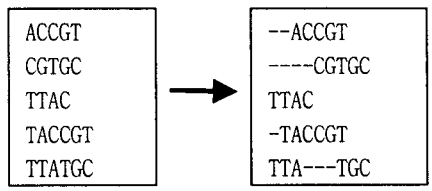

그림 4. 비정상적 염기 서열 단편(chimeric fragment)[4]
가장 아랫쪽에 있는 염기 서열 조각의 경우 서로 멀리 떨어진 두 염기 단편의 일부분이 하나로 합쳐진 것임을 알 수 있다.

이러한 잘못된 서열 조각들에 대한 고려 없이 일반적인 fragment assembly problem에 대한 알고리즘을 적용한다면 정체 불명의 생물체가 만들어질 것이다.

4. 몇 가지 해결 방안들

위에서 말한 Whole-Genome random shotgun approach에서 선결해야 할 몇 가지 문제에 대한 해결방안으로 각 염기 서열조각을 다른 것과 짝을 이루도록 하는 것이 있다. (그림5)

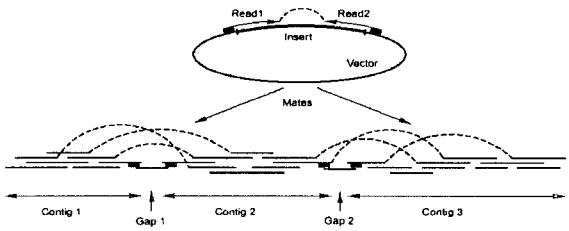

그림 5. mate[1]
cloning vector에 삽입된 유전자 조각(clone)의 염기 서열을 양쪽에서부터 읽어나가면 동일한 유전자 조각에 대한 염기 서열 분석 결과물이 두 개씩 짝을 이루게 된다.

(그림5)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 l_1, l_2라 했을 때 그 둘 사이의 거리는 약 $2000-(l_1+l_2)$ 혹은 $10000-(l_1+l_2)$ 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 콘티그를 만들었을 때, 짝을 이루는 두 개의 염기 서열 조각 사이의 거리가 계산치보다 더 짧다면 이는 잘못 구성된 콘티그라 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.

콘티그들 사이의 순서 정보도 짝을 이루는 두 염기 서열 조각을 통해 얻을 수 있다. Random shotgun 방법으로 유전체에 대한 염기 서열 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 콘티그들의 순서 정보가 매우 중요하게 쓰인다. (그림5)에서 볼 수 있듯이 콘티그들의 순서는 짝을 이루는 두 염기 서열 조각을 통해 얻어낼 수 있다. 두 콘티그 사이의 순서를 알아내는 방법으로는 이 외에도 공개 데이터베이스를 이용하여서도 알아낼 수 있다. 한 예로 GenBank를 들 수 있는데 이곳에는 현재까지 세계 각지에서 발견된 유전자들의 염기서열 혹은 아미노산 서열이 등록되어 있다. 이 GenBank 사이트에서는 임의의 염기 서열 혹은 아미노 서열을 입력하면, GenBank 내의 데이터 중에서 입력받은 것과 가장 비슷한 데이터(유전자나 특정부위의 염기서열)를 찾아 주는 서비스를 하고 있는데 만약 어떤 두 콘티그의 한쪽 끝 부분이 각기 같은 유전자(혹은 특정 부위의 염기서열)의 양쪽 끝 부분과 매우 유사하다고 밝혀지면 그 두 콘티그들은 서로 근접해 있다고 볼 수 있다(그림6).

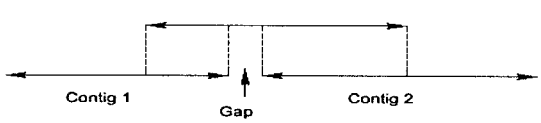

그림 6. 외부 데이터를 이용한 콘티그들 배열
콘티그1의 오른쪽 끝과 콘티그2의 왼쪽 끝이 각각 외부(ex, GenBank)에서 얻은 데이터의 왼쪽부분과 오른쪽 부분에

일치한다. 이를 통해 콘티그1과 콘티그2가 서로 인접한 콘티그라는 것을 알 수 있다.

이러한 방법들을 통해 콘티그들 사이의 순서를 정한 후 갭을 메울 수 있다. 콘티그들 사이의 순서를 정한 후에 갭을 메우는 작업은 여러 가지 방법으로 할 수 있는데 그 중에서 PCR기법을 이용하는 방법을 가장 많이 사용한다.

염기 서열 조각들의 방향에 대한 정보도 짝을 이루는 염기 서열 조각을 통해 얻을 수 있다. 다른 염기 서열 조각들과의 유사도를 통해 어떤 한 염기 서열 조각의 방향이 정해지면 자연히 그것과 짝을 이루는 염기 서열 조각의 방향도 정해진다. 직관적으로 이는 시간 복잡도를 반으로 감소시킴을 알 수 있다.

전체 염기 서열을 밝히는 작업을 힘들게 하는 요인 중의 하나인 서열 분석 에러는 random shotgun 방식을 이용할 경우 자동적으로 보정 가능하다. Random shotgun방식에서는 최소 6-fold 이상의 많은 수의 염기 서열 조각을 분석하므로 자체적으로 서열 분석 에러를 보정할 수 있다(그림7).

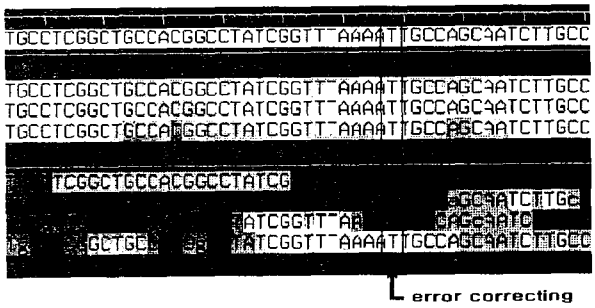

그림 7. 서열 분석 에러 보정
충분한 분량의 염기 서열 조각이 있을 경우 위 그림과 같이 잘못 분석된 염기를 보정할 수 있다. 빨간색 상자 안에 있는 다른 염기 서열은 모두 'T'이지만 아래에서 세 번째 줄의 그것은 'g'로 되어 있다. 따라서 그 부분의 실제 값(맨 윗줄)은 'T'로 인식, ('x'는 무시)

fragment assembly작업을 매우 어렵게 만드는 요인인 비 정상적인 염기 서열 조각(chimeric fragment)도 매우 많은 수의 염기 서열 조각을 분석함으로써 적절히 제거할 수 있다. 그렇지만 이런 경우 15-fold 이상의 아주 많은 수의 염기 서열 조각을 분석해야 하며 이는 매우 비용이 많이 드는 작업이다. 따라서 비 정상적 염기 서열 조각을 assembly작업 이전에 찾아내서 제거해야 한다. 비 정상적 염기 서열 조각을 찾아내기 위한 아이디어는 매우 간단하다. A라는 한 염기 서열 조각의 일부가 다른 염기 서열 조각에서도 발견된다면 A는 비 정상적인 염기 서열 조각이라고 할 수 있다. 이 분야는 매우 흥미로운 연구주제로, 좀 더 효과적으로 비 정상적 염기 서열 조각을 찾아내기 위한 연구가 현재 활발히 진행중이다[4].

5. 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. Celera Genomics는 whole-genome random shotgun 방식으로 *H.Influenzae*(약 1.8 M base pairs)와 *Drosophila Melanogaster*(약 120M base pairs)의 전체 염기 서열을 밝혀내어 random shotgun방식에 대한 노하우를 쌓은 후 인간 유전자에 대한 염기서열 분석작업을 시작한 것이다[1]. 이것이 시발점이 되어 각국에서 경쟁적으로 생명체의 염기서열을 밝히고 있으며, 본 저자들도 앞선 genome-project들에 대한 사전 연구 후에 실제로 미생물체의 염기 서열을 국내 최초로 분석하면서 random-shotgun방식에 대한 노하우를 쌓게 되었다. random shotgun방식은 결과적으로 HGP에 박차를 가한 것이 되었는데, 그 뒷편엔 '생물 정보학'이라 부르는 전산학적 뒷받침이 있었다. 따라서 HGP를 비롯한 다른 genome-project가 진행될수록 생물 정보학의 중요성은 더욱 커지게 될 것이다.

5. 참고 문헌

[1]. Gene Myers, "Whole-Genome DNA Sequencing"
[2]. J. Meidanis / J. C. Setubal, "Introduction to Computational Molecular Biology", pp.107-108.
[3]. E.S. Lander, M.S. Waterman, Genomics 2, 231(1988)
[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.
[6]. D. Gusfield, Algorithms on strings, trees, and sequences - Computer Science and Computational Biology, 1997.