# 메가 단위 미생물 유전체 규명을 위한 DNA 조각 재결합

마크로젠 정철희
고려대학교 최진영*
이화여자대학교 박현석

## 1. 서 론

DNA 조각 재결합 과정을 수반하게 되는 무작위 샷건 방법이 최근 몇 년 사이에 여러 게놈-프로젝트에 활발히 적용되면서 많은 생물체들의 전체 게놈 서열을 규명하는 데 크게 이바지 하였다. 수년 간 DNA 조각 재결합 작업 (혹은 DNA 어셈블리 작업)은 기술이 많이 향상 되었지만, 효율적이면서 정확한 최종 결과를 만들어 내는 DNA 조각 재결합 소프트웨어는 아직도 주요 연구 과제일 만큼 어려운 작업인 것이다. 2000년도 이미 97% 완성되었다고 발표한 휴먼 게놈이 3년이 지난 2003년에 와서야 100% 완성되었다고 발표된 것도 DNA 조각 재결합 작업의 어려움을 말해주는 좋은 예이다. 필자들 또한 Zymomonas mobilis의 전체 게놈 서열을 밝혀내는 과정에서 몇 가지 풀기 힘든 문제들을 경험하게 되었는데, 이러한 문제들은 DNA 조각 재결합 소프트웨어가 아직 여러 가지로 미흡하며 개선의 여지가 많다는 좋은 예로 볼 수 있다.

## 2. DNA 어셈블리 과정

전체적인 DNA 어셈블리 과정은 그림 1과 같이 요약할 수 있다.

이중 첫 단계인 샷건 데이터 준비과정은 DNA를 잘게 부수고, 다시 증폭시키기 위한 과정인데 다시 다음과 같은 단계를 거친다(3).

1. 초음파 혹은 튜브를 사용하여 분석하고자 하는 게놈을 잘게 조각낸다.
2. 전기 영동 등의 여러 가지 실험적인 기법을 통해서 일정한 길이의 DNA 조각(insert 혹은 클론)들을 선택한다.

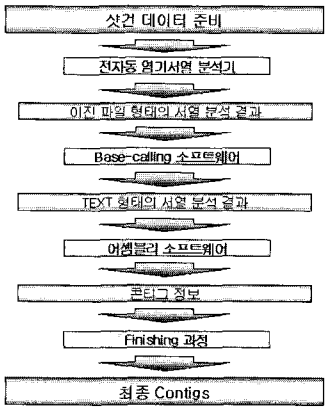



그림 1 DNA 조각 재결합 과정

3. 선택된 DNA 조각들을 대량 복제(클로닝)하기 위해 클로닝 벡터에 삽입한다.
4. DNA 조각이 삽입된 클로닝 벡터를 대장균 등의 미생물에 주입한 후 이를 대량으로 배양한다.

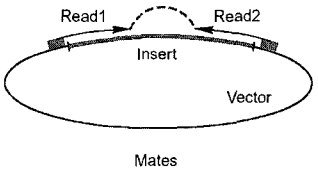



그림 2 메이트쌍 정보

* 종신회원

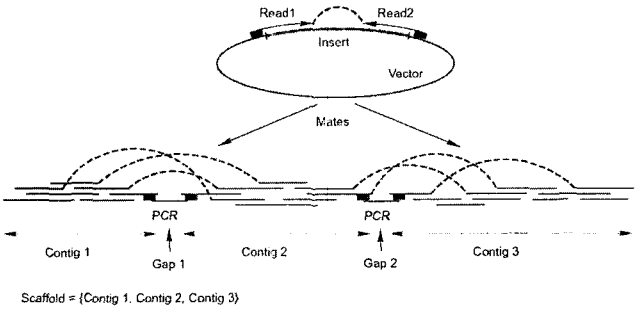



그림 3 메이트쌍 정보를 이용한 콘티그 정렬 및 갭 분석

이러한 과정을 거쳐 준비한 샷건 데이터의 염기서열 정보를 전자동 염기서열 분석기와 base-calling[1] 소프트웨어를 이용하여 읽어낸 것을 리드(read)라고 하는데, 이때 하나의 DNA 조각 양쪽 끝에서 두 개의 리드를 만들고 이를 메이트쌍(mate-pair)이라고 한다[오류! 참조 원본을 찾을 수 없습니다].

이렇게 얻은 리드들은 어셈블리 소프트웨어를 통해 어셈블리 과정을 거치면서 여러 개의 콘티그[2]로 만들어진다(그림 5 참조).

게놈에서 DNA 조각들을 추출할 때, 하나의 유전체를 놓고 샷건 처리하는 것이 아니라 여러 개의 동일한 유전체를 놓고 샷건 처리하기 때문에 각 DNA 조각이 만들어지는 것은 모두 독립적인 사건으로 간주할 수 있다. 다시 말해서 한 DNA 조각이 다른 DNA 조각의 발생 확률에 영향을 미치지 않는다. 따라서 DNA 조각의 발생 확률은 포아송 분포(Poisson distribution)을 따른다고 할 수 있다. DNA 조각의 발생 확률이 포아송 분포를 따르므로 실제 어셈블리 과정에서 사용되는 리드들 그 발생 확률이 포아송 분포를 따르게 된다. 그러므로 전체 게놈 상의 임의의 위치에 있는 염기가 리드들 중에 포함되어 있지 않을 확률을 다음과 같이 공식화할 수 있다[수식 1](11). 여러 개의 리드들이 모여서 연속적인 하나의 긴 염기서열을 이룬 중간단계 결과물

위 수식 1에 따르면 전체 리드 길이의 합이 게놈 전체 길이의 6배에 해당하면, 게놈 상의 임의의 염기가 리드들 중에 없을 확률은 $e^{(-6)}$이 되고 이 값은

수식 1 게놈 상의 임의의 위치에 있는 염기가 리드들 중에 포함되어 있지 않을 확률

$$P_0 = e^{(-\ell n/G)}$$

$P_0$ : 분석되지 않은 염기일 확률
$\ell$ : 리드의 평균 길이
G : 전체 염기서열의 길이
$n$ : 리드 수

0.0025이므로 이론적으로 리드들이 포함하지 못한 염기 서열은 거의 없어야 한다. 그러나 몇몇 생물학적인 요인들로 인해 실제로는 6배수에 해당하는 리드만으로는 충분치 않으므로, 10배수 이상의 리드를 준비해야 한다.

10배수 이상의 충분한 리드들이 준비되면, 어셈블리 소프트웨어들은 이들을 재조립해서 여러 개의 콘티그를 만들다. 물론 DNA 어셈블리의 궁극적인 목표는 완전한 하나의 콘티그를 만드는 것이지만, 충분한 리드를 준비한다고 하더라도 무작위 샷건 방법의 한계성 때문에 이런 결과는 거의 불가능하다(3). 따라서, 콘티그들을 원래 게놈과 최대한 일치하도록 배열한 후 추가적인 실험을 통해 콘티그들 사이의 갭(gap)들을 메우는 작업이 뒤따르게 된다. 이것이 최종 마무리 과정이고, 또 DNA 어셈블리 작업에서 가장 힘든 과정이다.

최종 마무리 과정의 핵심은 콘티그들 사이의 상대적인 순서 정보를 찾아내는 콘티그 정렬 작업인데 이 과정에서 앞서 설명했던 메이트쌍 정보가 매우 중요하게 사용된다. 메이트쌍 정보는 메이트를 이루는 두 리드들이 물리적으로 인접해 있다는 것과 이들 간의 거리는 DNA 조각의 크기와 같다는 사실에 기초하

1) DNA 조각의 염기(A/C/G/F)들의 배열 순서를 판별해 내는 것
2) 여러 개의 리드들이 모여서 연속적인 하나의 긴 염기서열을 이룬 중간단계 결과물

여, 여러 콘티그들 중에서 서로 인접한 두 콘티그를 판별해 낼 수 있고 둘 사이의 거리, 즉 갭의 크기도 추정할 수 있다. 이때, 메이트쌍 정보를 연속적으로 적용해서 콘티그들을 순서에 맞게 배열한 큰 덩어리를 스캐폴드(scaffold)라고 하며, 스캐폴드 내부에 있는 콘티그 사이의 갭들은 PCR등의 추가적인 실험을 통해서 분석한다[그림 3](3).

일반적으로 어셈블리를 위한 DNA 조각으로는 2k bp(base pair) 정도의 DNA 조각을 사용한다(1). 그러나 콘티그 배열 작업에는 10k bp나 30, 40k bp 혹은 100k bp 등 보다 긴 DNA 조각으로부터 얻은 메이트쌍 정보가 필요하다. 이런 추가적인 DNA 조각들을 클로닝할 때는 코스미드(COSMID), 포스미드(FOS- MID), 백(BAC) 등 약간 다른 종류의 클로닝 벡터를 사용한다.

콘티그 정렬 작업에 사용되는 또 다른 데이터는 분석할 게놈의 물리 지도(Physical map)이다. 물리지도란 게놈 상에 몇 가지 제한효소(restriction enzyme)[3] 가 인식하는 염기서열의 위치 및 몇몇 중요한 유전자들의 위치 등 콘티그의 실제 게놈상에서의 위치를 추정할 수 있을 만한 여러가지 단서들을 표시한 것이다. 이를 이용한 콘티그 정렬은 다음 그림 4와 같이 이루어진다.

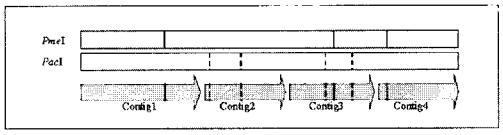



그림 4 제한효소의 위치정보를 이용한 콘티그 정렬

그림 4에서 위쪽 두 개는 각각 동일한 게놈에 제한효소 '*Pme* I'이 인식하는 염기서열의 위치와 제한효소 '*Pac* I'이 인식하는 염기서열 위치를 표시한 것이다. 이를 통해서 콘티그 1,2,3,4를 위의 그림 4와 같이 정렬할 수 있다. 게놈의 물리지도는 콘티그 정렬 과정뿐 아니라 정렬작업이 끝난 완성된 하나의 스캐폴드를 실제 게놈과 비교해서 어셈블리 및 콘티그 정렬 결과의 정확도를 확인할 때도 사용된다.

## 3. 리피트 서열과 메이트쌍

DNA 조각 재결합 과정에서 발생하는 여러가지 난제들 중 가장 해결하기 어려운 것은 반복적인 염기서열에 의해 발생하는 문제들이다. DNA 조각 재결합은 기본적으로 그림 5와 같이 스트링매칭 기법을 통해 이루어진다. 이때 사용되는 스트링매칭 기법으로는 스미스워터만(smith-waterman) 알고리즘(12)이라고 하는 국소정렬 알고리즘이다.

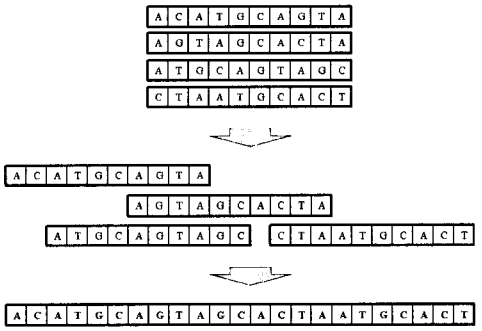


그림 5 스트링매칭을 이용한 리드들의 재결합 및 콘티그 생성

그러나 이 알고리즘을 따랐을 때, 리피트(repeat) 서열이 존재할 경우 스트링매칭 기법을 이용한 최종 DNA 결합 결과가 잘못될 가능성이 매우 높아진다[그림 6]. 예를 들어, 그림 6에서와 같이 Xc 부분이 재결합과정에서 한쪽으로 몰리게 된다.

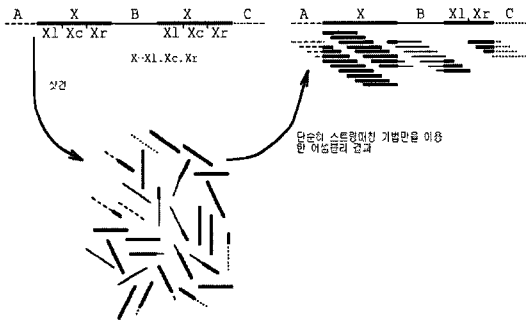



그림 6 리피트가 존재하는 DNA의 조각 재결합 작업

이와 같이 전체 DNA 염기서열 상에 리피트 영역이 존재할 경우, DNA 결합 과정에서 단순히 각 조각들의 유사도 정보만 사용한다면 잘못된 최종 결과가 나올 수 있다(4). 그러나, 실제로 유전체의 염기서열 중에는 여러 종류의 반복적인 리피트 영역들이 존재

3) 수 base-pair 길이의 특정 DNA 염기서열을 인식하여 DNA의 연결고리를 끊는 효소

하기 때문에 이러한 리피트 영역을 정확히 규명해 내기 위한 방법이 필요한데, 그 중 하나가 메이트쌍(Mate Pair) 정보를 이용하는 것이다.

메이트쌍은 앞에서도 언급했듯이 리드들 사이의 위치 정보에 대한 물리적인 증거를 제공하므로 DNA 조각 재결합 과정에서 발생하는 여러가지 난제들 중 특히 리피트에 의한 문제점을 해결할 수 있는 실마리가 된다. 다음절에서는 Zymomonas mobilis(이하 자이모모나스)의 예를 들어 가며, 실제 분석한 데이터를 제시하고, 그 과정을 좀더 자세히 설명해 보고자 한다.

## 4. 자이모모나스 데이터 분석 예

### 4.1 Zymomonas mobilis 데이터 준비

자이모모나스의 전체 염기서열 규명을 위한 샷건 데이터는 모두 3번에 걸쳐 만들어졌다. 이 중 주된 어셈블리에 사용된 데이터는 1차/2차 데이터로 모두 길이가 1000-2000 염기쌍 정도인 클론으로부터 나온 42,621개의 리드였고, 3차로 전체 정렬에 참조기 위하여 10k 데이터와 40k 포스미드(fosmid) 데이터를 각각 3천여개, 768개를 제작하였다. 이로써 총 리드들의 전체 DNA coverage는 약 12배로 추정되었다.

### 4.2 어셈블리 과정

1차/2차 데이터가 준비되어 감에 따라 PHRAP을 이용한 조각 재결합 작업도 동시에 진행하였다. 리드 수가 많아짐에 따라 콘티그 수도 줄어들고, 콘티그들의 총 크기도 2.4Mb정도에 수렴하는 등 외형적으로는 큰 문제 없이 진행되는 것처럼 보였다[그림 7].

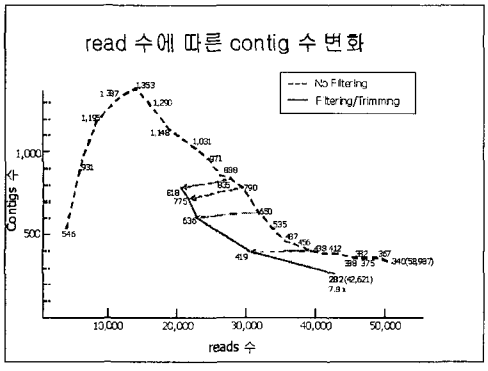



그림 7 1차/2차 데이터의 리드 수에 따른 콘티그 수의 변화

그러나 2차 데이터까지 제작이 끝난 후에도, 콘티그 수는 282개나 되고, 콘티그들 사이의 메이트쌍 정보도 매우 혼란스럽게 연결되어 있었다. 먼저 각 콘티그들 사이에 복잡하게 얽혀 있는 잘못된 메이트쌍 정보들을 제거하는 작업이 이루어져야 했다. 잘못된 메이트쌍 정보들이 엉켜 있는 경우 그 부분의 콘티그 서열이 정확하게 정렬 된 것이라고 보기도 힘들기 때문이다. 그래서, PHRAP의 결과 파일들에서 각 메이트쌍 들의 간격 또는 오프셋(offset) 정보를 추출한 후, 간격이 클론들 크기의 범위에 들어오는 데이터들만 사용하여 다시 436개의 콘티그를 제작하였다. 이 과정에서 콘티그 수가 많이 늘어나긴 했지만, 최종 결과에서 메이트 정보들이 어지럽게 얽혀있던 것이 거의 사라지는 등 전체적으로 매우 정돈된 결과를 보여주었다[그림 8].

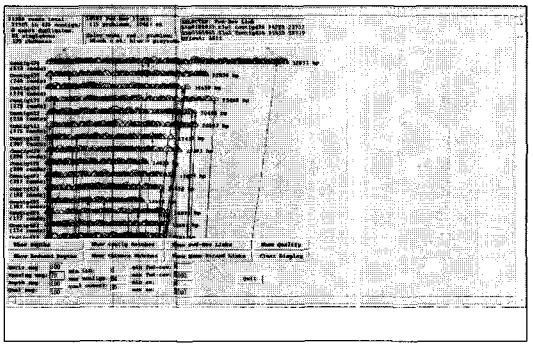


그림 8 436개 콘티그에서의 메이트쌍 정보

일단 메이트쌍 정보가 정돈된 어셈블리 결과를 얻은 후에는 3차 데이터에서 제작했던 약 2천개의 10k 데이터를 이용하여 개략적인 정렬 작업을 진행하였다. 이 때, 두 콘티그가 끝부분에서 서로 유사한 서열을 갖는다면 그 두 콘티그를 직접 연결하는 작업도 병행하였는데, 이러한 작업을 통해서 436개 콘티그를 8개의 임시 스캐폴드(scaffold)로 묶을 수 있었다.

그러나, 10k 데이터의 오프셋이 편차가 크게 나오고, 콘티그의 끝부분의 서열 유사도 만으로 직접 스캐폴드로 연결한 부분도 있는 등 8개 임시 스캐폴드 내부에 있는 서열이 오류 없이 결합된 것이라고 확신하기는 힘들었다. 그래서 이 8개의 스캐폴드에서 정확하게 결합되었다고 판단되는 부분들만 다시 잘라내는 작업이 필요했는데, 이때 메이트쌍 정보가 매우 중요하게 사용되었다.

DNA 조각 재결합 결과에서 메이트쌍 정보가 다음 그림 9와 같이 연속적으로 중첩되어 나타나는 곳은 곧 클론들 자체가 서로 중첩되어 있음을 의미하기 때문에 정확하게 결합된 곳이라 할 수 있다.

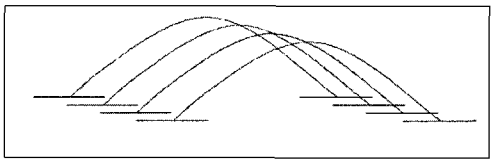


그림 9 메이트쌍 정보가 중첩되는 영역

따라서 8개 임시 스캐폴드에서 메이트쌍 정보가 위와 같이 중첩되어 나타나는 영역을 찾아 총 49개 최종 콘티그를 새로 제작할 수 있었다.

### 4.3 콘티그 정렬

49개의 최종 콘티그들의 정렬에는 크기가 약 40kb인 포스미드 데이터 768개(384 메이트쌍)와 실험에 의해 얻어진 Not I, Pac I, Pme I 이라고 하는 세 제한효소가 인식하는 위치를 비롯하여 20개 주요 유전자들의 위치를 표시한 물리지도를 사용하였다[그림 10](5).

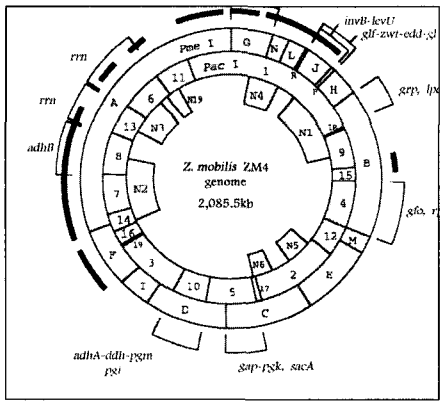



그림 10 Pac I, Pme I, Not I과 20개의 유전자를 이용한 물리지도

포스미드 데이터와 물리지도를 이용하여 49개 콘티그들을 12개의 갭(gap)이 없는 스캐폴드로 구성하고, 이를 기준으로 콘티그 정렬에 사용했던 제한지도를 다시 작성하였다[그림 11]. 이는 실험에 의해 작성된 그림 10과도 거의 일치하였다.

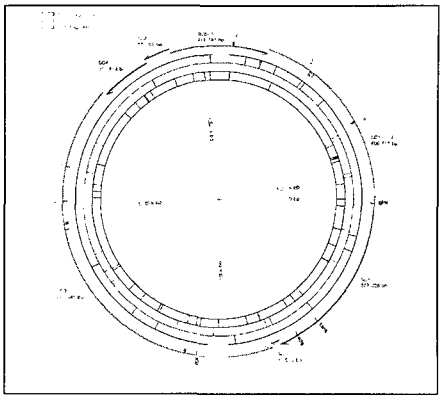


그림 11 자이모모나스의 새로운 물리지도

이후 추가적인 실험을 통해 갭을 메우는 작업을 성공적으로 끝마칠 수 있었다.

한편, 콘티그들을 정렬하는 과정에서 Ribosomal RNA(rRNA)라는 일종의 리피트에 의해 리드들이 잘못 결합되어 있는 부분을 발견할 수 있었다. rRNA는 길이가 약 5k bp 이며, 여러 미생물들 게놈 상에 2개 이상 존재하므로 일종의 리피트 영역이라 할 수 있는 것이다. 자이모모나스의 경우에도 rRNA가 세 개인 것으로 분석되었는데, 이 과정에서 rRNA에 의해 리드들이 잘못 결합된 부분을 수정해야 했다.

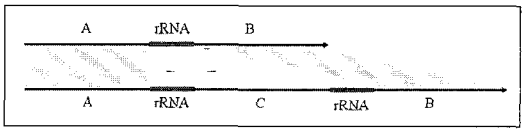



그림 12 rRNA에 의해 잘못 어셈블된 부분 수정

위의 그림 12에서 볼 수 있듯이, 위에 있는 콘티그는 rRNA에 의해 잘못 리드들이 잘못 결합되어 있었음을 알 수 있다. 원래의 염기서열은 두 rRNA 사이에 'C' 영역이 포함되어 있지만, 위의 콘티그에서는 첫 번째 rRNA와 두 번째 rRNA가 서로 합쳐지면서 'C' 영역이 빠진 채로 콘티그가 만들어진 것이다.

이렇게 하여 완성된 자이모모나스의 전체 게놈 서열의 총 길이는 2,056,416bp였다. 이 게놈 서열에는 3개의 ribosomal RNA(rRNA)가 존재했으며, 추정된 유전자 수는 2,112개였다. 이 중 1,249개는 이미 기능이 밝혀진 것들이고, 242개는 기능이 밝혀지진 않았지만 다른 생물체에서도 유전자일 것이라고 추정된

것이었으며, 나머지 621개 만이 자이모모나스의 게놈에서만 새롭게 발견된 것이었다.

## 5. 결 론

DNA 조각 어셈블리 과정은 스트링매칭에 기반을 두고 있기는 하지만, 바이오인포매틱스 분야에 전산학자들이 대거 참여하게 만들었으면서도 현재의 생물학 연구에 매우 큰 공헌을 하고 있는 것이기도 하다. 그러나, 수년 동안 매우 많은 DNA 조각 재결합 작업이 진행되었고, 여전히 많은 사람들이 더욱 우수한 DNA 조각 재결합 소프트웨어를 개발하려고 노력하지만, 생물학적 염기서열의 다양한 특성들로 쉽게 좋은 성과를 얻기가 힘들다. 그러나 이번 연구를 통해서 메이트쌍 정보가 정확한 어셈블리 결과를 만들어 내기 위한 매우 중요한 정보라는 것을 다시 확인할 수 있었다. PHRAP과 함께 지금까지 널리 사용되었던 몇몇 잘 알려진 어셈블리 소프트웨어들 중에도 어셈블리 과정에서 메이트쌍 정보를 활용할 수 있도록 만들어진 것들이 있긴 하지만, 이들보다는 PHRAP의 어셈블리 성능이 더 우수하다는 연구 결과가 있다(6). 그러나 최근 몇 년 사이에 개발된 어셈블리 소프트웨어들 중 Arachne(7)(8), Phusion assembler(9), EULER(10) 등과 같이 처음부터 메이트쌍 정보를 중점적으로 사용하도록 만들어진 것들은 이전 소프트웨어들보다 상당히 정확한 어셈블리 결과를 만들어 낸다.

이러한 어셈블리 소프트웨어들도 꾸준한 개선 작업을 통해 보다 정확한 어셈블리 결과를 만들어 내도록 하고 있는 등 어셈블리 작업을 위한 메이트쌍 정보 이용에 대한 연구는 계속되고 있다. 자이모모나스의 어셈블리 작업을 통해서 알게 된 DNA 조각 재결합 과정에 메이트쌍 정보를 적용하는 방법은 메이트쌍 정보를 이용하는 새로운 어셈블리 소프트웨어 제작은 물론 이미 메이트쌍 정보를 어셈블리 과정에 중점적으로 적용하고 있는 소프트웨어들을 개선하는 데에도 좋은 참고가 될 것이다.

## 참고문헌


[1] R.D. Fleischmann et al., "Whole-genome Random Sequencing and Assembly of H.influenzae," (1995) Science, Vol. 269, No. 5.223. 496-512.

[2] P. Green, "PHRAP documentation,"(1994). http://www.phrap.org.

[3] Eugene Myers, "Whole-genome DNA Sequencing," (1999), IEEE Computational Engineering and Science 3, 1, 33-43.

[4] Eugene Myers, "Toward Simplifying and Accurately Formulating Fragment Assembly," (1995), Journal of Computational Biology, 2(2), 275-290.

[5] Hyung-Lyun Kang and Hyen-Sam Kang, "A physical map of the genome of ethanol fermentative bacterium Zymomonas mobilis ZM4 and localization of genes on the map," (1998) Gene, 206(2), 223-228.

[6] Ting Chen and Steven S. Skiena, "A case study in genome-level fragment assembly," (2000) Bioinformatics 16, 494-500.

[7] Batzoglou, S., et al., "Arachne: A whole-genome shotgun assembler," (2002) Genome Res.12: 177.189.

[8] David B. Jaffe, et al., "Whole-genome sequence Assembly for Mamallian genomes: Arachne 2," (2003) Genome Res. 13: 91-96.

[9] Mullikin, J.C. and Ning, Z. "The Phusion Assembler," (2003) genome Res. 13: 81-90.

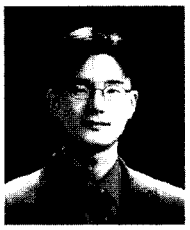
[10] Pavel A. Pevzner and Haixu Tang, "Fragment assembly with double-barreled data," (2001) Bioinformatics 17, 225-233.

[11] E.S. Lander and M.S. Waterman, (1988), Genomics 2, 231.

[12] T. F. Smith and M. S. Waterman, (1981) J. Mol. Biol. 147:195-197

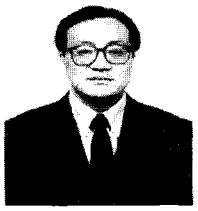


**정 철 희**



2000 고려대학교 이과대학 컴퓨터학과 학사
2002 고려대학교 이과대학 컴퓨터학과 석사
2000~현재 마크로젠 BI연구소 연구원
2002~현재 고려대학교 컴퓨터학과 박사과정
관심 분야 : Fragment assembly, Comparative Genomics, Gene Prediction

### 최 진 영



1982 서울대학교 컴퓨터공학과 졸업
1986 Drexel University Dept. of Mathematics and Computer Science 석사
1993 University of Pennsylvania Dept. of Computer and Information Science 박사
1993~1996 Research Associate, University of Pennsylvania
1994~1995 Computer Scientist, Computer Command and Control Company (Part time)
1996~1999 고려대학교 컴퓨터학과 조교수
1999~현재 고려대학교 컴퓨터학과 부교수
관심 분야 : 컴퓨터이론, 정형기법(정형 명세, Formal verification), 실시간 시스템, 분산 프로그래밍 언어, 소프트웨어 공학
E mail : choi@formal.korea.ac.kr

### 박 현 석

1986 서울대학교 전자공학과 졸업
1994 펜실베니아대학교 정보전산학 석사
1997 캠브리지대학교 컴퓨터학 박사
1997~1998 동경대학교 postdoctoral fellow
2000~2002 세종대학교 컴퓨터학과 조교수
1998년~현재 (주)마크로젠 이사
2003년~현재 이화여자대학교 컴퓨터학과 조교수
관심 분야 : 바이오인포매틱스, 텍스트 마이닝