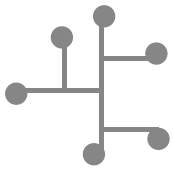


메타유전체 서열 조립의 문제점에 대한 연구

Issues on haplotype assembly of metagenomes

김우철(Woo-Cheol Kim)[1], 여윤구(Yun-Ku Yeo)[2], 김종현(Jong-Hyun Kim)[3], 박상현(Sang-Hyun Park)[4]

요 약

미생물의 경우 전체 생물종 중에서 많은 비율을 차지함에도 불구하고, 대부분의 미생물은 자연상태에서 분리시켜 배양하여 연구하기 힘들다. 최근의 미생물학은 자연상태에 있는 미생물을 직접 연구하려는 시도를 하고 있는데, 이런 연구를 가능하게 하기 위해서는 메타유전체의 염기서열을 밝혀내야 한다. 메타유전체의 염기서열을 밝혀내면, 현재 유전체학(Genomics)에 이용되는 분석방법을 적용할 수 있게 된다. 그러나, 단일 유전체가 아닌 특정 환경에서 살고 있는 다양한 미생물 유전체들의 집합인 메타유전체(Metagenome)의 경우에는 현재 확립된 서열조립 방법론이 없는 실정이다. 유전체의 염기 서열을 밝히는데 일반적으로 널리 쓰이고 있는 방법은 Whole Genome Shotgun Sequencing(WGSS)이다. 기존의 WGSS방식은 염기변이가 낮은 단일 유전체를 서열화(sequencing)하는데 초점을 맞춰서 개발되었기 때문에 염기변이율이 높은 유전체(polymorphic genome)의 서열화에 적용시킬 경우에는 염기서열의 연속성(continuity)이 심각하게 손상된다. 염기변이로 인한 염기서열 연속성의 손상은 여러 미생물들의 유전체서열이 동시에 조립되는 메타유전체의 경우에는 더욱 심각해 진다. 이와 같이 메타유전체를 대상으로 하는 서열조립에 대한 후속 연구들이 필요하다. 따라서 본 논문은 이러한 연구들의 기초 연구가 될 수 있는 메타유전체 조립 알고리즘에 대한 연구의 중요성과 메타유전체 서열화 과정에서 발생하는 문제점들의 원인을 분석한다. 이를 바탕으로 후속 연구들이 파생될 수 있는 기초 연구를 제공한다.

주제어 : 서열화, 유전체 조립, 메타유전체, 메타유전체학

이 논문은 2008년도 교육과학기술부의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(2008-2004103)

1 연세대학교 컴퓨터학과 박사과정
2 연세대학교 컴퓨터학과 석사과정
3 연세대학교 컴퓨터학과 연구교수
4 연세대학교 컴퓨터학과 부교수, 교신저자

+ 논문접수 : 2008년12월 30일, 심사완료 : 2009년 3월 25일

Abstract

Although microbial organisms take a significant portion of taxa, most of them are not trivially cultured under the laboratory condition. Recent advances in microbiology and availability of Metagenome sequences enable the study of microbial organisms under the natural condition. Although Metagenome sequences are essential to apply genomics technologies to the study of microbial organisms, the strategy to assemble Metagenome sequences is yet to be established.

Whole-genome Shotgun Sequencing (WGSS) has been widely used to sequence genomes. Because the assembly strategy of WGSS has focused on assembling less polymorphic genomes, it undermines the continuity of assembled genome sequences, applied to highly polymorphic genomes. The impairment of the sequence continuity becomes more serious in assembling metagenomes. In this paper, we note the problem of the impaired continuity, and emphasize the urgency of optimizing assembly strategy for metagenomes.

Keyword: sequencing, genome assembly, Metagenome, Metagenomes

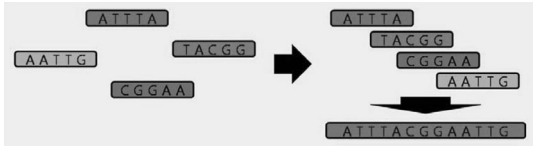
1. 서론

유전자(gene)는 생명체의 형질(phenotype)을 결정짓고, 유전을 통해 그 형질을 자손에게 전달하는 중요한 단위이다. 유전자는 생명체의 형질을 발현시키기 위한 기본적인 유전 정보로서 A, C, G, T 의 4가지 염기로 이루어졌다. 이러한 염기 정보는 mRNA로 전사되는 전사(transcription) 단계를 거친다. 그런 다음 전사된 mRNA로부터 단백질을 만들어 내는 번역(translation, 또는 해독) 단계를 거쳐 생명체의 형질을 발현한다. 따라서 유전자의 염기 배열에 따라 생명체의 형질이 결정되는 것이기 때문에, 생명체의 비밀을 풀기 위해서 유전자의 염기 서열을 알아내는 과정(sequencing)은 필수적이다.

하지만 특정한 기능을 담당하는 유전자가 유전체(genome)의 어느 부분에 있는지 개별적으로 밝혀 내는 것은 비효율적이다. 따라서 일차적으로 각각의 유전자가 아닌 생명체의 유전체 정보 전체를 서열화한 다음 유전자를 예측해 나가는 것이 현재 유전체학(Genomics)의 접근방법이다[1]. 그 한 예가 인간의 유전체 염기 서열을 서열화한 Human Genome Project[3][4]이다. 이렇게 서열화된 인간 유전체 정보를 바탕으로 질병에 관여하는 염기변이를 밝혀내는 연구나 반복되어서 나타나는 염기 서열인 CNV(Copy Number Variation)를 찾는 연구 등 많은 후속 연구들이 진행되고 있다[5].

유전체의 크기는 생명체에 따라서 편차가 있다. 박테리아와 같은 미생물의 DNA경우에는 수백만 bp(base pair)의 염기로 구성되어 있다. 인간을 포함한 포유류 동물의 DNA는 수십억 bp의 염기로 구성되어 있다. 그러나 현재의 DNA 염기 서열 판독 기술은 한번에 서열화 가능한 DNA 조각의 길이가 최대 500-800bp에 불과 할 뿐만 아니라 특정 위치부터 판독하는 것 역시 불가능 하다[6]. 이런 기술적 한계 때문에 현재의 DNA 염기 서열 판독 기술만으로는 유전체 전체를 한 번에 서열화할 수 없다.

이런 기술적인 한계를 극복하면서 유전체 서열을 분석하기 위한 방법으로 널리 사용되는 기법이 Whole Genome Shotgun Sequencing(WGSS)이다. WGSS는 다음과 같은 단계로 구성되어 있다. 먼저 서열화 하려는 유전체에서 전체 DNA 크기의 7~8배 정도의 샘플을 채취하여 잘게 부순다. 이 때 채취하는 샘플의 배수를 서열 중첩 배율(sequence coverage)라고 한다. 그 다음, 염기 서열 판독 기술을 이용해서 각각의 조각을 서열화한다. 마지막으로 그림 1과 같이 조각들의 서열에서 서로 중첩되는 부분을 연결하는 과정을 반복적으로 적용하면서 전체 유전체를 조립한다.



〈그림 1〉 유전체 조각을 중첩되는 부분을 이용하여 다시 조립하는 과정

지금까지 초파리나 인간의 유전체와 같은 많은 생물의 유전체가 이러한 연구방식을 통해 서열화되었다. 그렇지만 자연에 존재하는 생명체의 다양성에 비하면 그 숫자는 아직 일부에 불과하다. 한 통계학적 분석에 따르면, 이론적인 관점에서 살펴볼 때 전체 해양에는 약 2백만 종, 토양에는 1톤당 약 4백만 종의 미생물이 존재하는 것으로 알려져 있다[7]. 현재 유전체가 완전히 밝혀져 공개된 생물이 NCBI 데이터베이스를 기준으로 3,194종에 불과한 것을 감안하면 아직도 유전체를 밝혀야 할 생물이 다수이다.

그러나 기존의 연구방식만으로는 자연계의 모든 생물을 연구하기 어렵다. 가장 대표적인 원인이 난배양미생물(VBNC, Visible But Non-Culturable)의 존재이다.

난배양미생물이란 기존의 순수배양 환경에서는 배양되지 않는 미생물을 말한다. 숙주를 벗어나면 생존이 불가능한 병원균이나 공생 생물 또한 이런 범주에 속한다. 이렇게 순수배양이 불가능한 생물의 경우 기존의 방식으로는 유전체 조립에 필요한 샘플을 충분히 확보하기가 어렵게 된다.

또한 순수배양을 통해 확보한 샘플만을 이용하여 유전체를 조립하는 경우, 자연 상태에서 존재하는 생명체의 다양성에 대한 정보를 얻을 수 없게 된다. 대부분의 미생물은 환경의 영향이나 생명체간의 상호작용의 영향을 받으면서 생명활동을 유지한다. 이 과정에서 환경에 적응하기 위해 미생물의 형질이 일부 변화하기도 하는데, 기존의 연구방식은 이런 생명체의 다양성을 반영하지 못한다. 기존의 순수배양기술은 서식 환경에서 하나의 생물만을 분리하여 인위적인 환경에서 배양하는 작업이기 때문이다.

이런 문제점을 해결하기 위하여 최근 부각되고 있는 연구분야로 메타유전체학(Metagenomics)이 있다. 메타유전체학은 생명체 하나의 유전체가 아니라 어떤 환경에 존재하는 생물 집합 전체의 유전체를 연구 대상으로 하는 학문이다. 메타유전체학은 연구실 환경이 아닌, 생명체가 서식하는 환경에서 직접 샘플을 채취하기 때문에, 배양이 불가능한 미생물뿐 아니라 서식 환경에 존재하는 생명체간의 상호작용에 대한 정보까지 얻을 수 있다.

메타유전체학을 통해 어떤 환경에서 살고 있는 생명체들의 유전체를 분석하는 경우, 자연 환경에서 서식하는 생명체의 종류와 양이 매우 불규칙하다. 그러나 현재의 서열화 알고리즘은 대부분 한 생물체의 염기 서열이 있는 경우만을 가정하고 개발된 알고리즘이기 때문에, 다양한 생물의 집합에 이를 그대로 적용할 경우 좋은 결과를 얻어내기 힘들다. 따라서 본 논문에서는 메타유전체 조립 알고리즘에 대한 연구의 중요성과 메타유전체 서열화 과정에서 발생하는 문제점들의 원인을 규명하고자 한다.

본 논문에서는 2장에서 WGSS에 대해서 논의한 후 3장에서 메타유전체학에 대해서 소개한다. 4장에서는 메타유전체의 서열 조립의 기초 연구로 다염기변이 유전체 조립에 관련된 내용을 소개한다. 5장에서는 메타유전체학에 대한 소개와 함께 서열화 과정에서 다루어야 하는 중요한 논점에 대해서 논의한다. 마지막으로 6장에서 결론을 기술한다.

2. WGSS : Whole Genome Shotgun Sequencing

대개의 WGSS 연구에서는 유전체 정보를 분석하기 위해서 다음과 같은 단계를 거친다. 첫 단계는 먼저 유전체를 조각 내어 라이브러리(library)를 구축하는 단계이다. 다음 단계는 유전체 조각을 서열화한 뒤 중첩(overlap)을 기반으로 조립(assembly)하는 단계이며, 마지막 단계는 조립된 조각들의 순서와 방향, 연결 관계를 조합하여 최종적으로 일치된 서열(consensus sequence)를 만드는 단계이다.

2. 1. 라이브러리 구축

WGSS에서는 먼저 유전체를 초음파나 고압의 가스를 이용해 작은 조각으로 부순다. 그 다음에 유전체 복제체(clone vector)라고 불리는 외부 DNA를 받아들일 수 있는 물질에 유전체 복제 조각(clone insert)을 삽입한 뒤, 이를 배양하여 유전자 양을 증폭한다. 이런 과정을 통해 유전체 복제 라이브러리(clone library)가 구축된다. 만약 유전체 샘플 하나만을 조각 내어 라이브러리를 구축하였다면 당연히 중첩된 영역이 하나도 없게 되

며, 원본 유전체를 다시 조립할 수가 없게 된다. 따라서 충분한 중첩 영역을 갖기 위해서 모든 유전체 복제 라이브러리는 여러 개의 유전체 샘플을 이용한다. 각각의 라이브러리는 얼마나 많은 샘플을 사용하여 라이브러리를 구축했는지를 서열 수집 배율(clone coverage)로 표시하고 있으며, 대개 20배(20-fold) 이상의 값을 갖고 있다. 20배란 라이브러리를 구성하고 있는 유전체 복제물(샘플)이 20개라는 것을 의미한다.

이후 라이브러리의 각 유전체 복제 조각의 양쪽 끝부터 각각 염기 서열을 읽어내 한 쌍의 염기 서열을 만들어 낸다. 염기 서열 판독에 사용되는 방법은 F.Sanger와 A.R.coulson이 개발한 Dideoxy chain termination (또는 Sanger dideoxy sequencing) 방법[6]과 A.Maxam과 W.Gilbert의 Chemical degradation 방법[8]이 있다. 현재는 서열화 장비(sequencing machine)에서 구현이 용이하다는 장점 때문에 Dideoxy chain termination 방법을 선호하고 있다.

Read들을 조립해 다시 유전체를 복원해 내려면 충분한 중첩 영역을 확보해야 한다. 그렇기 때문에 유전체 조립에 사용할 read를 만들어 낼 때, 원본 유전체의 길이보다 더 많은 양의 read를 만들어 내야 한다. 이론적으로 read가 원본 유전체의 99% 이상을 해독하려면, 유전체 복제 배율(sequence coverage)이 5배 이상이 될 때까지 read를 만들어야 한다[9]. 예를 들어, 2Mbp의 박테리아 유전체를 조립할 때에는 약 10Mbp(read 하나의 길이가 700bp라고 가정할 때 약 15,000개)의 read를 채취해야 원본 유전체의 99% 이상을 해독할 수 있다.

염기 서열을 판독하는 방법들은 각 read의 염기 서열 정보와 함께 read의 각 염기 (A, T, G, C)의 판별 정확도를 나타내는 정확도(quality score)를 함께 제공한다. 이 값은 양의 정수로 값이 높을수록 염기의 판독이 정확함을 나타낸다. 이 정확도 q는 식 1.을 이용해서 확률값 p로 바꿀 수 있다. 예를 들어 정확도가 20이라면, 이 위치의 염기가 정확하게 읽어졌을 확률은 $1-10^{-2}$으로 99%가 된다. 이러한 정확도 값은 서열의 조립 과정에서 중요한 정보로 사용된다.

$$p = 1 - 10^{-\frac{q}{10}}$$

〈식 1〉 염기 서열 판독 정확도 q를 확률값 p로 변경하는 식

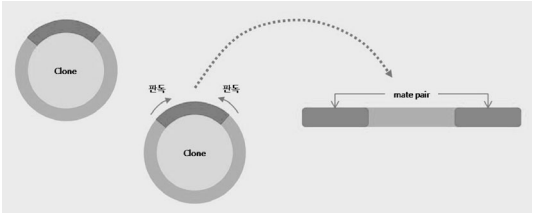
그림 2와 같이, 같은 유전체 복제 조각에서 읽어낸 두 개의 sequencing read들을 mate pair라고 한다. Mate pair로부터 얻을 수 있는 가장 중요한 정보는 같은 라이브러리에서 만들어낸 read들을 포함하는 유전체 복제 조각의 길이와 read의 길이가 일정하다는 것이다. 이런 정보를 이용하면 mate pair인 read 사이의 거리를 추정할 수 있기 때문에 이러한 정보는 조립 과정에서 중요한 정보로 사용된다. 다른 정보는 read를 읽어낼 때 유전체 복제 조각의 양쪽 끝에서부터 읽어내기 때문에 mate pair인 두 read의 방향(orientation)은 서로 반대라는 것이다.

일반적인 유전체 조립 연구에서는 유전체 복제 조각의 길이가 8~10Kb정도의 길이가 긴 라이브러리와 2~3Kb정도의 길이가 짧은 라이브러리를 각각 하나 이상씩 포함한다. Celera genomics에 의해서 완성된 Human Genome Project의 경우 2Kb, 10Kb, 50Kb의 길이를 가지는 3개의 라이브러리를 이용했으며, 유전체 복제 배율은 5.3배였다[20].

2.2. 조립 단계

WGSS의 다음 단계는 조각난 read들을 조립하여 원래의 유전체를 복원해 내는 것이다. 조립 알고리즘의 가장 기본적인 전략은 read 사이의 중첩 영역을 찾아낸

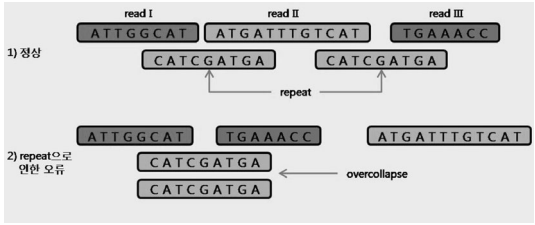
다음, 해당 read를 연결하는 방법이다. 하지만, 이러한 단순한 전략만 사용하는 경우 유전체의 조립의 정확도가 낮을 수밖에 없다. 그 이유는 다음과 같다. 이 방법은 중첩되는 부분이 있는 두 read는 유전체의 같은 부분에서 얻어진 것이라고 가정하고 있다. 그러나 유전체에는 일정한 염기 서열이 규칙적으로 반복되는 영역인 repeat가 존재한다. 이 repeat가 있기 때문에 서로 중첩되는 두 read가 반드시 유전체의 같은 부분이라고 확정할 수가 없게 된다. 따라서 이러한 단순한 전략에 기반한 조립은 정확도가 낮을 수 밖에 없다.



〈그림 2〉 유전체 복제 조각을 삽입하고 양쪽 끝에서부터 판독하여 한 쌍의 sequencing reads를 만든다.

이러한 repeat은 생명체 안에 다수 존재하기 때문에, - 인간의 경우, 약 10만 번이나 반복해서 나타나는 repeat 영역이 존재한다[3]. - repeat은 유전체의 서열화를 어렵게 하는 중요한 원인이다. 예를 들어, 그림 3와 같이 3개의 read 사이를 repeat 2개가 연결하고 있는 것이 원래의 유전체인 경우, repeat을 잘못 조립하게 되면 read I과 read III이 repeat에 의해 연결되고 read II를 별도의 조각으로 판단하는 결과가 나타날 수 있다. 또한 이 경우 repeat 2개가 같은 위치에 중복 위치되어 있는데, 이것을 과중첩(overcollapse) 현상이라고 한다. 다른 위치에 있어야 할 read가 한 위치에 중복되어 위치되기 때문에 이로 인해 원래 연결되어야 할 read들이 따로 떨어져 조립되게 된다.

이러한 이유 때문에, repeat을 탐지하고 그것에 대응



〈그림 3〉 repeat으로 인해 유전체가 잘못 조립되는 상황

하는 것은 WGSS 알고리즘이 풀어야 하는 중요한 문제점 중 하나이다. 올바른 조립 결과를 얻기 위해서는 조립 과정에서 항상 repeat을 고려하면서 read를 조립해 나가야 한다. 조립 작업이 끝난 이후에 repeat을 탐지하고 잘못된 조립 결과를 수정하는 것은 굉장히 많은 노력과 비용을 요구하는 작업이기 때문이다. 따라서 repeat에 대응하는 것은 조립 단계에서 수행되어야 하며, 그것은 조립 알고리즘의 성능을 판단하는 중요한 기준이다.

단순하지만 가장 일반적으로 사용되는 repeat 탐지 방법은 통계적 수치를 활용하는 것이다. 원본 유전체에서 read가 완전히 무작위(random)로 채취되었다고 가정할 때, read를 정렬한 결과가 대체로 균등하게 분포해야 한다는 것이 이 방법의 주요 아이디어이다. 만약 어느 한 부분이 다른 부분에 비해 훨씬 많은 read가 정렬되어 있다면 그 부분은 repeat으로 인한 과중첩 부분일 확률이 높다고 간주한다. 물론 이 방법 역시 완벽하지 않다. 유전체 복제물을 이용해 read를 만드는 과정에서 특정 부분의 유전체 복제 배율이 다른 부분에 비해 높거나 작을 가능성이 있으며, 작은 크기의 repeat일 경우에는 탐지할 수 없다는 약점도 있다. 하지만, 이런 통계적인 방법은 repeat을 탐지하는 일반적인 기준으로 사용된다.

조립 단계에서 repeat을 탐지하기 위해 사용할 수 있는 또 다른 전략은, 먼저 repeat으로 의심되는 read들은 별도로 분리해 놓고, repeat이 확실히 아닌 read들

을 기준(anchor)으로 삼아 그 read에서부터 중첩된 read들을 연결해 중첩된 read들의 집합인 contig를 확장해 나가는 방법이다. TIGR Assembler[19]가 이런 방법을 사용하였다.

조립 과정에서 repeat을 해결하는 다른 접근방식으로 그래프에 기반한 방식은 Eulerian path를 이용한 연구가 있다[10]. 먼저 각 read를 서로 중첩되는 n-mers로 분할한 뒤, 각각의 edge가 n-mers를 나타내는 그래프를 그린다. 이 때 각 edge에 연결된 양 끝 node는 n-1 접두어와 n+1 접미어가 된다. 이렇게 read의 조립을 그래프 방식으로 생각하면, 조립 문제는 결국 모든 edge를 한번씩 포함하는 path인 Eulerian path를 찾는 것이 된다. 이렇게 read를 연결해 나가면 과중첩이 생기는 것을 근본적으로 방지할 수 있다. 이런 이론상의 장점에도 불구하고, 아직까지 큰 크기를 가진 실제 유전체의 서열화에서 사용된 적이 없고 대부분의 유전체 조립 방법들은 Eulerian path기반이 아닌 overlap-layout-consensus 기반의 알고리즘을 사용하고 있다.

2.3. 마무리 단계

Read를 contig로 조립하는 조립 작업이 끝나더라도, 그 결과물이 하나의 contig로 나타나는 경우는 거의 없다. 유전체 중첩 배율이 7~8 정도에 불과할 뿐 아니라, read를 라이브러리에서 무작위로 채취하기 때문에, 전체 유전체 중에서 염기 서열이 판독된 read가 없는 끊어진 부분이 존재하게 된다. 따라서 여러 개의 contig를 하나로 조립하여 완성된 유전체를 만드는 과정이 필요하게 된다. 이런 작업을 scaffolding이라고 하며, 이 작업에서는 contig의 방향과 순서를 맞추어 더 큰 염기 서열인 scaffold로 조립하게 된다. 현재 대부분의 서열 조립 알고리즘은 scaffolding 과정을 포함하고 있다.

Scaffolding 작업에서는 mate pair의 정보가 중요하게 사용된다. 만약 서로 다른 두 개의 contig에 mate pair를 이루는 read가 각각 하나씩 포함되어 있다면, 두 contig는 인접해 있는 것으로 추정할 수 있다. 또한, mate pair인 두 read의 방향은 서로 반대이어야 하기 때문에, 이 방향 정보를 이용해서 contig의 방향을 정할 수 있다. Contig들을 조립해 본 결과에서 mate pair인 두 read의 거리가 라이브러리의 유전체 복제 조각의 길이에 비해 너무 가깝거나 멀다면 유전체 조립 과정에 오류가 있었음을 의미한다. 또한, mate pair의 방향이 서로 같은 방향이라면 이것 역시 유전체 조립 오류로 판단할 수 있다.

Contig를 scaffold로 조립하는 과정은 그래프로 표현할 수 있다. 각각의 contig가 node가 되고 mate pair로 연결된 contig간의 연결 여부가 edge가 된다. 이렇게 scaffold를 찾는 과정을 그래프 알고리즘으로 바꾸어 보면, scaffolding의 최종 목표는 그래프를 하나의 선이나 원으로 만들어 내는 작업이 된다. (일부 미생물의 경우 genome이 원형이다.)

Scaffolding 과정을 거치더라도 그 결과가 반드시 하나의 scaffold로 나타나는 것은 아니다. 특정 영역의 낮은 유전체 복제 배율, 염기 서열 해독 과정에서 에러, repeat으로 인한 잘못된 염색체 조립 등의 원인으로 여러 개의 scaffold가 산출되거나, scaffold 안에 gap이 생겨날 수 있다. 이 중 scaffold 내부에 존재하는 gap을 sequence gap이라고 하는데, 이것은 walking 테크닉이라고 하는 기법으로 쉽게 그 gap을 메울 수 있다[11]. 다른 scaffold 사이에 존재하는 gap은 physical gap이라고 하는데, 이것을 처리하려면 많은 양의 수작업과 실험적인 방법을 필요로 한다.

2.4. 서열 조립의 평가

이렇게 gap을 메운 다음에 최종 산출물인 일치된 서열을 찾을 수 있다. 이러한 일치된 서열은 유전체의 연속적인 염기서열이 여러 개의 scaffold를 포함하게 된다. 각각의 scaffold의 길이는 수십 Mb부터 수 kb까지 매우 다양할 수 있다. 서로 다른 scaffold는 원칙적으로 전체 유전체상에 중첩되지 않으며, 다른 scaffold사이에는 우리가 알 수 없는 작은 길이의 공간이 존재한다. 유전체 조립 알고리즘의 결과물은 밝혀진 염기서열도 정확해야 하지만 산출된 scaffold의 크기가 충분히 커야 한다.

서열 조립 결과의 우수성을 평가할 때 scaffold의 크기가 충분히 크다면 scaffold가 연속성 측면에서 우수하다고 할 수 있다. 가장 이상적인 경우는 하나의 scaffold가 하나의 염색체인 경우이다. 하지만 일반적인 서열 조립에 사용되는 서열 중첩 배율이 7~8에 불과하다는 것을 고려하면 이러한 경우는 불가능하다. Scaffold들이 연속성 측면에서 우수하지 않으면, 서열 조립 결과의 의미가 없기 때문에 많은 문제점이 생길 수 있다. 가장 대표적인 것은 scaffold길이가 짧아지면 유전자를 예측하기가 힘들어진다는 것이다.

흔히 조립된 염기서열의 연속성을 평가하는 단위로 N50 scaffold length를 사용한다. N50 scaffold length란 조립된 scaffold를 길이에 따라 큰 순서부터 작은 순서로 배치할 때, scaffold길이의 총합의 중간에 위치하는 scaffold의 길이이다. 염기변이가 낮은 생물의 유전체 조립에 관련되어서는 현재 많은 진전이 있어 왔고, 유전체 조립의 연속성 측면에서도 우수하다. 현재 인간 유전체의 N50 scaffold length는 2.7 Mb이고 침팬지 유전체의 경우에는 2.3 Mb정도 되는 것으로 알려져 있다[20][21].

3. 메타유전체학

기존의 유전체 연구는 연구실 환경에서 순수배양이 가능한 생물들을 대상으로 하고 있다. 그러나 순수배양으로 배양 가능한 미생물은 전체의 1%에 불과하며, 자연계에서 중요한 역할을 수행하는 미생물의 대부분이 난배양 미생물인 것으로 추정되고 있다[12]. 또, 순수배양이 가능한 미생물들 조차도 유사한 환경에서 배양 가능한 생물들이기 때문에 결국 비슷한 성질을 갖고 있는 것이 대부분이다[2]. 이로 인해 유전체의 다양성을 확보하는 데에 제약을 받고 있으며, 이러한 한계점을 극복하기 위해 부각되고 있는 학문 분야가 바로 메타유전체학이다.

메타유전체학에서는 연구실 환경에서 배양된 단일 유전체가 아니라, 어떤 환경(Environment)에서 유전체를 한꺼번에 채취하여 그것을 연구의 대상으로 삼는다. 유전체를 채취하는 환경은 토양, 바닷물, 배설물, 화석과 같이 수많은 미생물이 공존하고 있는 서식지를 대상으로 한다.

메타유전체에서 유전체를 분석하는 작업은 전체적으로 단일 유전체를 분석하는 작업과 유사한 부분이 많다. 먼저 샘플로부터 DNA를 채취한 뒤, 여러 가지 생물의 DNA가 섞인 상태에서 한꺼번에 DNA를 잘게 부순다. 이어서 라이브러리를 구축하고, 염기서열 해독과정을 거쳐서 유전체 서열화에 필요한 read를 획득한다. 이후 조립 알고리즘을 이용해서 DNA 조각을 다시 조립한다. 마지막으로, 조립된 DNA를 Blast등의 유틸리티를 이용하여 기존에 구축된 DB정보와 비교하여 환경에 서식하고 있는 생물의 계통을 판별하거나 새로운 유전자에 대한 정보를 확보할 수 있다.

유전체 분석 과정의 마지막 단계에서 조립된 DNA를 기존에 구축된 DB정보와 비교하는 것은, 메타유전체에는 수많은 생물의 유전체가 섞여 있기 때문이다. 그 때

문에 메타유전체를 다시 조립해 내도, 조립된 염기 서열이 어느 생물의 유전체에 속하는 지를 구분하기가 어렵다.

이런 문제점 때문에, 현재까지의 메타유전체학에서는 전체 유전체의 서열을 완전히 밝히는 것이 아니라, 다양한 유전자를 분석하고 연구하는 방향에 초점을 맞추고 있다. 이런 유전자 중심 접근법도 유전자 정보를 확보한다는 점에서 생물학적으로 가치를 지니고 있다. 한 예로 특정 서식 환경에서 지배적으로 나타나는 유전자를 의미하는 EGT(Environmental Gene Tags)가 있다. 만약 어떤 서식 지역에서는 유전자 A가 다량 발견되고, 다른 서식 지역에서는 유전자 B가 다량 발견된다면 유전자 A와 B가 해당 서식지에서의 생존과 관련된 유전자임을 추정할 수 있다. 이러한 연구를 통해 유전자의 기능을 밝혀낼 수 있으며, 이렇게 밝혀낸 유전자 정보를 이용하여 난배양 미생물의 순수배양에 성공한 사례도 있다[13].

또한 메타유전체학을 통해 환경에 서식하는 미생물의 종류를 확인하고 새로운 미생물을 발견할 수 있다. 모든 세포 생물은 16S rRNA라고 하는 비교적 작은 염기 서열을 갖고 있다. 이것은 각 계통에 따라 고유한 서열을 갖고 있기 때문에 세포 생물의 계통을 확인하는 지표로 사용되고 있다[14]. 메타유전체학을 통해 얻은 16s rRNA를 기존의 DB와 비교·분석함으로써 해당 환경 내에 서식하는 미생물의 종류를 확인할 수 있으며, 새로운 미생물을 발견할 수도 있다. 폐광 유출수(Acid mine drainage)에서 발견되는 바이오 필름(biofilm)을 분석한 연구[15]에서는 shotgun sequencing 방법을 이용하여 3종의 거의 완전한 유전체와 2종의 부분적인 유전체를 복원해 냈다. 이런 방법으로 다양한 환경에서 채취한 샘플을 비교 분석하게 되면, 각각의 환경에 따른 생물군의 구성 변화를 연구할 수 있다. 한 예로, Susannah Green Tringe등의 연구[16]에서는 바닷물(Sargasso sea), 심해의 고래 사체(Whale fall), 농장의 토양(Minnesota farm soil), 폐광 유출수(Acid mine drainage)의 데이터를 이용하여 생물군의 구성을 비교하였다.

또한 위에서 예로 들었던 EGT와 같이 어떤 환경에서의 유전자 분포를 분석하면, 서로 다른 환경에서 미생물이 어떻게 적응해 나가고 있는지를 연구할 수 있다. 예를 들어, 북서 대서양과 태평양의 바닷물을 분석한 연구[16]에서는 북대서양과 태평양에 서식하는 미생물의 대략적인 분류와 수를 비교하였다. 총 41개의 서로 다른 위치에서 바닷물 샘플을 채취했으며, 샘플과 샘플 사이는 약 320Km 정도의 거리를 두고 결정되었다. 분석 결과 파나마 운하 근해와 뉴저지 연안(북대서양), 갈라파고스 군도(태평양)가 각각 뚜렷하게 구분되는 유전자 분포를 보였는데, 이 세 가지의 분류군은 수온, 바닷물의 깊이 등 서식지의 특성이 서로 구분되는 집단이었다. 이런 결과는 서식 환경에 따라 생물이 다른 유전자 변화를 겪으며 적응해 왔다는 것을 의미한다. 이 연구를 통해서 발견된 1,700여개의 단백질 군(cluster)들이 이미 알려진 단백질 패밀리(family)와 상당한 차이를 보인다는 것을 발견하였다. 이것은 메타유전체학을 통해 앞으로 많은 양의 새로운 단백질을 발견할 수 있다는 것을 보여주는 예가 될 수 있다. 또 심해에 있는 고래의 사체(Whale fall)를 분석한 연구[16]에서는 서로 수천 Km가 떨어진 고래의 사체를 샘플로 사용하였지만, 분석 결과 유사한 유전자 구성을 확인함으로써, 비슷한 환경에서 서식하는 생물은 유사한 유전자 구성을 가진다는 것을 확인하였다.

또한 메타유전체 연구는 인간의 질병 연구에 기여하는 바가 크다. 인간의 몸 속에는 많은 미생물이 서식한다고 알려져 있고, 장에만 10^{13}~10^{14} 종의 미생물들이 서식한다고 추정된다. 인간의 몸 속에 살고 있는 미생물들을 총칭하여 Human Microbiome이라고 부르고,

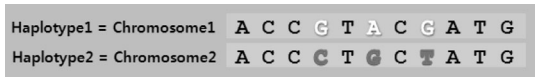
Human Microbiome은 인간의 유전자의 수보다 최소 100배 많은 유전자를 가지고 있을 것이라 예측된다. 이 경우 각각의 인간 개개인은 Human Microbiome에 다른 환경을 제공하고 있고, Human Microbiome은 다른 환경에 적응하며 진화하고 있는 것이다. 인간의 몸 속에 살고 있는 미생물들과 인간의 상호관계를 밝히는 것은 인간의 질병연구에 핵심적인 단서를 제공할 수 있다. 메타유전체적 분석 방법을 통해 이런 관계를 밝히는 소규모 연구[24]가 최근 발표되었고, 5년간 $115백만 불이 투입된 The Human Microbiome Project[1]라는 대규모 프로젝트가 시작되었다[31].

최근에 유전체 서열화 연구에 관련되어 두 가지 흥미로운 방향이 제시되었다. 첫째, 서열화에 들어가는 비용을 줄이기 위해 대용량 서열화(Ultra-throughput sequencing)를 추구하고 있다. 앞에서 설명한 pyrosequencing이 그 대표적인 예가 될 수 있다. 미생물들의 유전체는 보다 복잡한 생물체, 예를 들어 인간이나 침팬지에 비해 repeat가 전체 유전체에서 차지하는 비중이 적고 유전체의 크기도 작다. 따라서 대용량 서열화 기법을 적용하기 적합한 면을 가지고 있다. 메타유전체 연구에서 대용량 서열화 기법이 적용되어 가고 있다. 메타유전체로부터 일배체형을 추출하는 방법은 현재 개발되어 있지 않으나, 이런 방법론이 제안될 경우 조립된 메타유전체로부터 특정 환경에 존재하는 미생물들의 종이나 population structure를 보다 쉽게 예측할 수 있다.

둘째, 조립된 유전체 서열로부터 일배체형을 유추해 내는 것이다. 현재 멍게(*Ciona intestinalis*)로부터 일배체형을 추출하는 방법론이 제안되었으며[27][28], 인간 한 명의 유전체로부터 일배체형을 추출하는 작업이 완료되었다[30]. 이런 일배체형의 추출은 생물들간의 유전적인 관계를 밝혀내고, 인간의 질병 연구에 큰 도움을 줄 수 있다는 것이 밝혀졌다. 또한 이러한 연구들은 메타유전체를 위한 서열화 연구의 기초 연구들로 사용될 수 있다.

4. 메타유전체 서열 조립을 위한 기초 연구: 다염기변이 유전체 서열 조립

두 개의 상동 염색체(Homologous chromosomes)가 존재하는 생물일 경우 각각의 염색체들은 일배체형(haplotype)으로 정의 된다. 그림 4의 예와 같이 상동 염색체는 Chromosome1과 Chromosome2의 2개의 일배체형으로 구성되어 있고 일부 염기의 경우 일배체형마다 서로 다른 염기로 나타나는 염기변이(Polymorphism)를 포함하고 있다.



〈그림 4〉 상동 염색체의 예 : 서로 다른 색으로 표시된 G-C, A-G, C-T는 염기변이를 나타냄.

염기서열을 밝히려고 하는 유전체에 염기변이가 많을 경우에는 두 개의 염색체의 서열은 상당히 차이가 난다. 이렇게 염기변이가 많은 유전체의 경우에는 하나의 유전체만을 대상으로 하는 기존의 방법으로 조립해 나가기 힘들다. 멍게 (*Ciona intestinalis*), 성게 (*Strongylocentrotus purpuratus*)같은 생물들의 유전체 서열을 밝히기 위한 지놈 프로젝트를 수행하면서 이런 문제를 해결하려고 했으나 기존의 방법에서 벗어나지 못해서 scaffold의 연속성 측면에서 좋지 못한 성과를 얻었다[18].

멍게 (*Ciona intestinalis*)의 염기변이의 수를 유전체의 크기로 나눈 비율인 염기변이율은 1.2%로 알려져 있고, 유전체서열의 N50 scaffold length는 190 kb에 불과하다[18]. 성게 (*Strongylocentrotus purpuratus*)의

1 http://nihroadmap.nih.gov/hmp

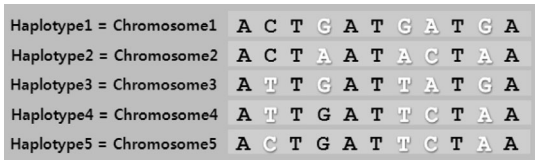
염기변이율은 4~5%로 알려져 있고, N50 scaffold length는 142 kb에 불과하다[21]. 멍게(*Ciona intestinalis*)와 성게(*Strongylocentrotus purpuratus*) 유전체 서열조립에는 염기변이율이 낮은 유전체 서열조립에 쓰이던 방법들이 쓰였다.

다염기변이 유전체의 서열을 조립하기 위한 방법을 개발해 또 다른 멍게 (*Ciona savignyi*)의 유전체 서열 조립에 쓰였지만 이 방법은 서열 중첩 배율을 통상적인 것보다 두 배 가까이 요구하는 약점이 있었다[22]. 다른 멍게 유전체 조립의 N50 scaffold length는 989 kb이었지만 서열 중첩 배율이 통상적인 것보다 두 배에 육박하는 ~13을 사용하였다[22]. 이렇게 서열 중첩 배율을 두 배 가까이 올릴 경우에 지놈 프로젝트를 수행하는데 소요되는 비용도 2배 가까이 증가하게 된다. 이런 다염기변이 생물들의 N50 scaffold length는 인간과 침팬지의 N50 scaffold length가 2.7 Mb와 2.3 Mb라는 것을 생각할 때 염기서열의 연속성이 염기변이로 인해 얼마나 손상되어 있는 지 알 수 있다 (인간과 침팬지 염기변이율은 ~0.1%이다) [20][21].

다염기변이 유전체 서열 조립의 근본적인 문제는 두 개의 다른 일배체형에서 유래한 유전체 조각들이 기존의 유전체 서열 조립 방법을 사용할 경우 서로 다른 scaffold로 조립되는 경향이 있다는 것이다. 또한 이러한 경향은 염기변이율(polymorphism rate)이 높을수록 뚜렷해진다. 현재 지놈 프로젝트를 수행하는 컨소시엄들은 이 문제를 기존의 서열 조립 방법을 그대로 사용하거나 기존의 서열 조립 방법을 약간 개선해서 해결하려고 하고 있다[20][22][23].

5. 메타유전체 서열 조립의 문제점

다염기변이 유전체서열을 조립하는 문제도 현재 좋은 해결책이 없는 상태이지만, 이 문제는 염기변이가 결국에는 하나의 개체로부터 온 것에 한정된다. 하지만 메타유전체는 특정 장소의 흙이나 바닷물 혹은 인체의 소화기관으로부터 채취한 샘플에 존재하는 수많은 미생물로부터 얻어진 유전체들의 총합이기 때문에 염기변이로부터 야기된 서열조립 문제의 복잡성은 훨씬 심각해진다 (그림 5). 이러한 문제를 해결하기 위해서 최근 3년 사이에 선구적인 연구들 [17][14][15][24][25][16][26]이 진행되었다. 하지만 이러한 선구적인 연구들 역시 문제의 근본적인 해결을 하지 못해서 염기변이가 많지 않은 유전체를 조립할 때 사용하는 것을 계속해서 사용하고 있는 실정이다. 따라서 본 논문에서는 메타유전체 서열 조립의 문제를 근본적으로 해결하기 위해서 메타유전체 서열 조립의 문제점들을 그 원인들과 함께 나열한다.



〈그림 5〉 메타유전체에서의 일배체형: 다염기변이 유전체 서열 조립문제에서와 다르게 두 개 이상의 일배체형이 존재할 수 있음.

문제점 1: 유전체 구성의 복잡성

현재까지의 전유전체 조립 알고리즘들은 염기변이(polymorphism)가 낮은 유전체의 서열화 (예: 포유류의 유전체)에 있어서 좋은 성과를 올려왔다. 하지만 미생물과 같이 염기변이가 높은 생물에 있어서는 조립 결과의 연속성 면에서 저조한 성능을 보이고 있다. 예를 들어 인간의 염기변이율(염기변이의 수를 유전체의 크기로 나눈 비율)은 0.1% 정도이며, 유전체서열의 N50 scaffold length는 2.7MB이다. 반면 염기변이가 큰 생물인 멍게(*Ciona intestinalis*)의 경우 염기변이율은

1.2%로 알려져 있으며, N50 scaffold length는 190 kb에 불과하다[19]. 이런 현상은 read들을 합쳐서 contig를 만들 때 원래는 연결되어야 할 read가 염기변이로 인해 연결되지 않았기 때문에 발생한다.

이런 문제점을 해결하기 위해서는, read들을 합쳐서 contig를 만들어 나갈 때 어떤 identity level을 기준으로 하는지가 관건이다. 만약 그 기준이 너무 엄격하면 같은 contig로 조립해야 할 read들이 다른 contig들로 조립이 되어버리고, 반대로 너무 유연하면 다른 contig로 조립되어야 할 read들이 같은 contig들로 조립되어 버리는 결과를 낳는다. 염기변이가 적을 경우에는 이런 identity level을 결정하는데 별 문제가 없지만 염기변이가 많은, 혹은 여러 유전체들을 동시에 조립할 경우에는 이 level을 결정하는 것이 쉬운 일이 아니다. 단일 생물의 유전체를 서열화할 때는 예측된 유전체 크기를 가지고 이런 level을 조절해 나갈 수 있는 여지가 있지만, 메타유전체와 같이 여러 유전체들을 동시에 서열화할 경우는 문제는 더욱 복잡해진다. 메타유전체의 서열조립에 있어서 서열의 연속성 저하가 심각함에도, 이 문제를 해결하려는 시도는 아직 미약한 실정이다.

문제점 2: 이종성(Heterogeneity)으로 인한 조립의 어려움

또한 genomic rearrangement와 같은 현상으로 인한 유전체들간의 이종성은 유전체 조립을 어렵게 만드는 요인이다. Genomic rearrangement란 염색체가 분할되었다가 다시 합쳐지는 과정에서 발생할 수 있는 현상이다. 이 과정에서 DNA의 결손(deletion), 중복(duplication), 혹은 순서가 뒤바뀌는 현상(inversion)이 나타날 수 있기 때문에, 조립 결과에 영향을 미칠 수 있다. 이런 문제는 메타유전체 구성의 복잡성으로 인해 단일 유전체를 서열화할 때와 마찬가지로 복잡할 수 있다. 중복이나 순서가 뒤바뀌는 현상으로 야기된 서열조립의 문제점은 결국 repeat를 어떤 식으로 조립할 것인가로 귀결된다.

문제점 3: repeat 조립의 어려움

한 종의 생물을 대상으로 하는 유전체 조립 알고리즘의 경우, 어떤 부분의 유전체 복제 배율이 유난히 높으면 그 부분을 repeat으로 간주하는 통계학적 방법을 사용할 수 있다. 그렇지만 메타유전체를 대상으로 하는 유전체 조립 알고리즘은 그와 같은 통계학적 알고리즘을 사용할 수가 없다. 왜냐하면 생명체가 살고 있는 환경 내 존재하는 생물의 종마다의 개체 수 차이가 유전체 복제 배율에 반영될 수 있기 때문이다. 환경 내에 많이 존재하는 생물의 경우 유전체 복제 배율이 높을 것이고, 적게 존재하는 생물의 경우 유전체 복제 배율이 매우 낮을 것이다. 만약 단일 유전체를 조립하는 알고리즘을 그대로 사용한다면, 유전체 복제 배율이 높은 영역이 그 생물의 고유한 염기 서열임에도 불구하고 그것을 repeat으로 처리할 위험이 있다.

Sargasso sea의 연구에서는 특히 유전체 복제 배율이 너무 높은 영역을 repeat으로 처리되는 문제점을 해결하기 위해 단계적으로(tiered fashion) 유전체를 조립하는 방식을 채택했다. 먼저 최초의 조립 단계에서 1) 크기가 크고 2)높은 유전체 복제 배율을 가지면서 3)반복적이지 않은 contig를 수작업으로 탐색한다. 이 때 탐색한 contig들의 데이터를 바탕으로 최종적인 조립 결과물에서 고유한 영역(unique region)이 가질 유전체 복제 배율을 추정한다. 그 이후부터는 단계적으로 조립을 진행하면서, 매 단계마다 최초 단계에서 추정했던 값을 바탕으로 좋은 coverage를 가진 영역을 별도로 분리해 낸다. 이런 과정을 거치기 때문에, 기존의 조립 알고리즘에서는 repeat으로 처리되었을 영역을 고유한 염기 서열로 구분할 수 있다.

문제점 4: Library 구축의 어려움

메타유전체를 서열화에 함에 있어서 library 구축 측면의 두 가지 문제점이 있을 수 있다.

첫째, 특정한 환경에서 메타유전체를 서열화할 때에, 그 환경을 대표하는 미생물들이 서열조립에 적정 수준으로 clone library화 되어야 한다. 즉 많은 양의 DNA 조각들을 library화하고 서열화할수록 많은 미생물들의 유전체들이 메타유전체 서열에 포함될 수 있다. 이럴 경우에 많은 양의 DNA를 서열화해야 하기 때문에 비용 상승이 필수적이다. 둘째, library를 만들 때에는 편향성(bias)이 존재할 수 있다. 따라서 특정 생물들은 의도하지 않게 배제되어 메타유전체 서열에 포함되지 않을 가능성이 존재한다.

이에 대한 해결책으로 흔히 제안되는 것이 pyrosequenicing이다[28]. Pyrosequencing은 library 구축 단계를 거치지 않기 때문에, 편향성에 따른 문제가 없고, 서열화 비용이 훨씬 적게 든다. 하지만, 한번에 서열화할 수 있는 길이(300~500bp)가 짧고, 정확성이 Dideoxy chain termination보다 떨어진다는 단점이 있다.

6. 결론

메타유전체학은 현재 유전체 서열화의 도움에 힘입어 한 개별 유전자의 관점이 아니라 유전체의 관점으로 변화하고 있다. 이런 변화는 인간 유전체학이 인간 지놈 프로젝트를 통해 개별 유전자의 관점에서 전체 유전체의 관점으로 변화하는 것과 유사하다고 볼 수 있다. 인간 지놈 프로젝트를 완성시키기 위해서 많은 연구와 방법론의 개선이 이루어졌듯이 새로운 분야인 메타유전체학의 발전을 위해서 메타유전체학에 적합하게 유전체 서열조립 방법론의 개선이 요구되는 실정이다.

하지만 아직까지는 이 분야에 대한 기초 연구가 활발하지 못하다. 따라서 본 논문에서는 이 분야에 대한 연구의 방향에 대해서 소개하고 및 흥미로운 연구 주제들을 제안하였다. 구체적으로 살펴보면 먼저 본 논문에서는 메타유전체 조립 알고리즘을 개발 하기 위해 필요한 3가지 중요 내용에 대해서 소개하였다. 1) 먼저 일반적인 유전체 서열 조립에 관련된 기초 연구들을 소개하고 2) 메타유전체학에 대한 소개와 함께 메타유전체 조립 알고리즘에 대한 연구의 중요성을 소개하였다. 3) 마지막으로 기존의 유전체 서열 조립 방식과 메타유전체 서열 조립 방식의 차이점들과 함께 메타유전체 서열과 과정에서 발생할 수 있는 문제점들의 원인을 규명하였다.

본 논문에서 소개한 기초 연구들을 바탕으로 메타유전체를 위한 서열 조립 알고리즘을 개발할 경우, 메타유전체학이 목표로 하는 다른 환경에 적응하는 미생물들의 진화과정을 유추해 나가는데 큰 도움을 줄 것이다.

7. 감사의 글

이 논문은 2008년도 교육과학기술부의 재원으로 한국과학재단의 지원을 받아 수행된 연구입니다(2008-2004103).

8. 참고문헌

[1] E. W. Myers et al., “A Whole-genome Assembly of Drosophila”, Science, Vol. 287, pp 2196-2204, 2000.

[2] S. G. Tringe et el., “Metagenomics: DNA Sequencing of Environmental samples”, Nature Reviews Genetics, Vol. 6, No. 11, pp 805-814,

2005

[3] E. S. Lander et al, "Initial sequencing and analysis of the human genome", Nature, Vol. 409, No. 6822, pp 860-921, 2001.

[4] J. C. Venter et al, "The Sequence of the Human Genome", Science, Vol. 291, pp 1304-1351, 2001.

[5] R. Redon et al, "Global variation in copy number in the human genome", Nature, Vol. 444, pp 444~454, 2006.

[6] F. Sanger et al, "DNA sequencing with chain-terminating inhibitors", PNAS, Vol. 74, pp 5463-5467, 1977.

[7] T. P. Curtis et al, "Estimating prokaryotic diversity and its limits", PNAS, Vol. 99, pp 10494-10499, 2002.

[8] A. M. Maxam and W. Gilbert, "Sequencing end-labeled DNA with base-specific chemical cleavages", Methods Enzymol, Vol. 65, pp 499-560, 1980.

[9] E. S. Lander and M. S. Waterman, "Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis", Genomic, Vol. 2, pp 231-239, 1988.

[10] P. A. Pevzner et al., "An Eulerian path approach to DNA fragment assembly," PNAS, Vol. 98, pp 9748-9753, 2001.

[11] S. Batzoglou et al., "ARACHNE: A Whole-Genome Shotgun Assembler", Genome Research, Vol. 12, pp 177-189, 2002.

[12] P. Renesto et al., "Genome-based design of a cell-free culture medium for Tropheryma hipplei", Lancet, Vol. 362, pp 447?449, 2003.

[13] C. R Woese et al., "Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya", PNAS, Vol. 87, pp 4576-4579, 1990.

[14] W. Tyson et al., "Community structure and metabolism through reconstruction of microbial genomes from the environment", Nature, Vol. 428, pp 37-43, 2004.

[15] S. G. Tringe et al., "Comparative Metagenomics of Microbial Communities", Science, Vol. 308, pp 554-557, 2005.

[16] D. B. Rusch et al., "The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific", PLOS Biology, Vol. 5, No. 3, pp 398-431, 2007.

[17] J. C. Venter et al., "Environmental Genome Shotgun Sequencing of the Sargasso Sea", Science, Vol. 304, pp 66-74, 2004.

[18] P. Dehal et al., "The Draft Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins", Science, Vol. 298, pp 2157-2167, 2002.

[19] G. Sutton et al., "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects", Genome Science & Technology, Vol. 1, No. 1, pp 9-19, 1995.

[20] S. Istrail et al., "Whole-genome shotgun assembly and comparison of human genome assemblies", PNAS, Vol. 101, No. 7, pp 1916-1921, 2004.

[21] Chimpanzee Sequencing and Analysis Consortium, Nature 437, 69-87, 2005.

[22] Vinson et al. 2005

[23] The Sea Urchin Genome Sequencing Con-

sortium 2006.

[24] S. R. Gill et al., "Metagenomic Analysis of the Human Distal Gut Microbiome", Science, Vol. 312, No. 5778, pp 1355-1359, 2006.

[25] N. Kannan et al., "Structural and Functional Diversity of the Microbial Kinome", PLOS Biology, Vol. 5, Issue 3, pp 467-478, 2007.

[26] S. Yooseph et al., "The Sorcerrer II Global Ocean Sampling Expedition: Expanding the Universe of Protein families", PLOS Biology, Vol. 5, Issue 3, pp 432-466, 2007.

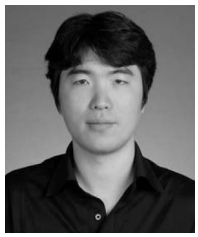
[27] J. H. Kim, M. S. Waterman, L. M. Li, "Diploid genome reconstruction of Ciona intestinalis and comparative analysis with Ciona savignyi", Genome Research, Vol. 17, Issue 6, pp 1101-1110, 2007.

[28] M. Margulies et al., "Genome sequencing in microfabricated high-density picolitre reactors", Nature, Vol. 437, pp 376-380, 2005.

[29] L. M. Li, J. H. Kim, M. S. Waterman, "Haplotype reconstruction from SNP alignment", Journal of Computational Biology, Vol. 11, pp 505-516, 2007.

[30] S. Levy et al. "The Diploid Genome Sequence of an Individual Human", PLOS Biology, Vol. 5. Issue 10, pp 2113-2144, 2007.

[31] P. J. Turnbaugh, "The Human Microbiome Project", Nature, Vol. 449, pp 804-810, 2007.



김우철

2003.8: 연세대학교 컴퓨터과학과(공학사)

2006.2: 연세대학교 컴퓨터과학과(공학석사)

2006.3~현재: 연세대학교 컴퓨터과학과(박사과정)

관심분야: 데이터베이스, 바이오인포메틱스, 유사검색

E-mail: twelvepp@cs.yonsei.ac.kr

여윤구

2009.3: 연세대학교 컴퓨터과학과(공학사)

2009.3~현재: 연세대학교 컴퓨터과학과(석사과정)

관심 분야: 바이오인포매틱스, 데이터베이스, 데이터마이닝

E-mail: yyk@cs.yonsei.ac.kr

김종현

2001.2: 연세대학교 경영학과 (경영학사)

2006.5: University of Southern California 컴퓨터과학과 (공학박사)

2007.3~2007.7: 연세대학교 컴퓨터과학과 박사후 연구원

2007.7~2008.5: University of Pennsylvania 의대 Postdoctoral Researcher

2008.6~현재: 연세대학교 컴퓨터과학과 연구교수

관심분야: 바이오인포매틱스

E-mail: jonghkim@cs.yonsei.ac.kr

박상현

1989.2: 서울대학교 컴퓨터공학과(공학사)

1991.2: 서울대학교 컴퓨터공학과(공학석사)

2001.3: UCLA대학교 전산학과(공학박사)

2001.2~2002.6: IBM T.J Watson Research Center Post-Doctoral Fellow

2002.8~2003.8: 포항공과대학교 컴퓨터공학과 조교수

2003.9~2006.8: 연세대학교 컴퓨터과학과 조교수

2006.9~현재: 연세대학교 컴퓨터과학과 부교수

관심 분야: 데이터베이스, 데이터마이닝, 바이오인포메틱스, 적응적 저장장치 시스템

E-mail : sanghyun@cs.yonsei.ac.kr