다염기변이 유전체에 대한 서열 정렬 툴 분석

김유선○ 김종현 여윤구 김우철 박상현
연세대학교 컴퓨터과학과
yoursun@cs.yonsei.ac.kr, angangdori@gmail.com, {yyk, twelvepp, sanghyun}@cs.yonsei.ac.kr

Analysis of sequence alignment Tools on polymorphic genomes

YooSun Kim○, Jong Hyun Kim, Yun Ku Yeo, Woo-Cheol Kim, Sanghyun Park
Department of Computer Science, Yonsei University

요 약

생명공학 기술의 발달로 지놈 프로젝트를 통해 인간・초파리 등 여러 종의 유전체 정보가 밝혀 졌다. 그러나 Post-Genome 연구에 있어서 매우 중요한 생물체인 멍게(*Ciona intestinalis*)와 성게(*Strongylocentrotus purpuratus*)의 유전체 서열은 현재 공개되어 있으나 염기서열의 연속성(continuity)에는 심각한 문제점이 존재하고 있다. 이들은 염기서열에 변이가 많은 다염기변이 유전체(polymorphic genomes)로 그 특성이 반영되지 않은 전통적인 Whole Genome Shotgun Sequencing(WGSS)방법을 사용였기 때문이다. 이와 같은 다염기변이 유전체 서열 분석은 시스템 생물학이나 비교 유전체학 등의 후발 연구에 기초가 되므로 매우 중요하다. 본 논문에서는 다염기변이 유전체에 대해 알아보고 서열 조립 알고리즘의 기본이 되는 서열 정렬 툴들 중 가장 많이 사용되는 FASTA, BLAST, BLAT에 대해 분석하여 봄으로써 다염기변이 유전체에 적합한 서열 조립 전략 수립을 위해 고려해야 하는 사항들을 논의해 본다.

1. 서 론

유전자(gene)는 유전정보의 기본단위로 A(아데닌)・C(시토신)・G(구아닌)・T(티민)으로 표시되는 4가지 염기의 배열로 이루어져 있다. 유전자는 중심원리(central dogma)에 의해 전사(trascription)와 번역(translation)을 거쳐 생명체의 형질을 발현하게 되며, 유전을 통해 그 형질을 자손에게 전달하게 하는 중요한 물질이다. 결국 유전자의 염기 배열에 의해 생명체의 형질이 결정되는 것이기 때문에 유전자의 염기서열 분석(sequnecing)은 생명체 연구에 있어 기초적이고 필수적인 작업이라 할 수 있다.

현재 유전체학(genomics)은 유전체(genome) 서열 분석을 위해 한 생명체의 유전체 정보 전체를 서열화한 다음 유전자를 예측해 나가는 접근 방법을 취하고 있는데, 이를 위해 유전체 서열 조립(sequence assmbly) 알고리즘이 사용되고 있다. 생명공학의 대용량처리(high-throughput) 기술 발달에 힘입어 유전체 조립 알고리즘에도 대용량 처리기술을 이용한 Whole Genome Shotgun Sequencing(WGSS) 방법이 개발되었는데, 이는 유전체를 무작위로 자른 작은 조각들을 염기서열 판독 기술을 이용하여 서열화하고, 이 서열에서 서로 중첩되는 부분을 연결하면서 유전체를 조립하는 방법이다[6]. WGSS 알고리즘들은 서열 정렬 툴(alignment tools)의 알고리즘에 기반한 서열 조립 전략을 세우고 있으며, 조립 결과에 대한 분석에도 정렬 툴들을 이용하고 있다. 그러므로 서열 조립 전략을 세우는데 있어 서열 정렬 툴을 이해하는 것은 필수적이다.

† 본 연구는 교육과학기술부 과학재단의 특정연구개발사업(2007-03965)의 지원을 받아 수행되었습니다.

현재까지 진행되어 온 인간이나 초파리와 같이 염기변이가 적은 유전체들의 분석은 기존의 WGSS 방법들에 의해 쉽게 조립되어 서열화된다. 그러나 우리가 살고 있는 자연계에는 수많은 미생물들을 포함하여 염기변이(sequence polymorphism)가 많은 생명체들이 훨씬 많이 존재하고 있으며, 이런 다염기변이 유전체(polymorphic genomes) 서열 분석은 post-genome연구에서 중요한 화두가 되고 있다. 대표적으로 다염기변이 유전체의 서열을 비교하므로써 비교유전체학 (comparative genomics)나 진화생물학 (evolutionary biology)을 연구하는데 기초가 되며, 미생물들이 생산하는 항생물질을 발견해 내어 신약 개발 및 질병연구 등에 기여할 수 있다. 그런데 멍게들(*Ciona intestinalis*, *Ciona savignyi*)과 성게(*Strongylocentrotus purpuratus*)와 같은 다염기변이 유전체들의 염기서열을 조립하는 데에 전통적인 WGSS 방법을 사용했을 때에 한계가 있다는 것이 밝혀졌다[8]. 또한 현재 생물정보학(bioinfomatics)의 흐름이 시스템생물학(system biology)과 같은 기능적인 방향을 지향하고 있기 때문에, 시스템생물학에서 중요한 다염기변이 생물들의 염기서열 분석의 문제점은 이러한 후발 연구의 장애요인으로 작용할 수밖에 없다. 다시 말하면 수많은 다염기변이 유전체들의 염기서열이 결정되어 가야만 이를 이용한 다른 연구들의 효율적인 진행이 가능할 것이므로 다염기변이 유전체들의 특성을 반영한 효과적인 유전체 조립 방법의 개발이 필요하다.

본 논문에서는 이와같은 다염기변이 유전체 조립 전략을 세우기 위해 기본이 되는 기존의 서열 정렬 툴들을 분석하고 이들을 다염기변이 유전체에 적용했을 때의 문제점에 대해 논의하고자 한다. 이를 위해 먼저 다염기변이 유전체와 전통적 WGSS 방법에 대해 간략히 살펴보고, 서열 정렬 툴의 알고리즘을 분석하여 이들의 특성을

알아본다. 또한 각 서열 정렬 툴들에 대해 다염기변이 유전체인 멍게들(*Ciona intestinalis*, *Ciona savignyi*)과 성게(*Strongylocentrotus purpuratus*)의 서열 정렬 수행 시간을 비교하여 봄으로써 다염기변이 유전체 조립 알고리즘에 있어 고려해야 할 점들에 대해 정리해 본다.

2. 다염기변이 유전체와 전통적 WGSS 방법

2.1 다염기변이 유전체

다염기변이 유전체는 genomic rearrangement 과정 등에서 DNA의 결손(deletion), 중복(duplication), 순서 뒤바뀜 현상 등을 통해 염기서열에 변이가 많이 있는 유전체를 말한다.

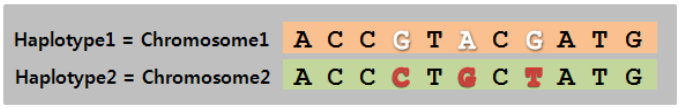

그림 1. 두 개의 상동 염색체 (homologous chromosomes)가 존재하는 다염기변이 생물에서 두 개의 일배체형(haploid)들에서 다른 부분들

일배체 다염기변이 유전체의 경우는 각 개체간에만 염기변이가 나타나게 되지만, 두 개의 상동 염색체를 가지는 이배체 다염기변이 유전체의 경우는 (그림 1)에서와 같이 두 개의 일배체형(haploid)간에도 많은 차이를 가지게 된다. 염기변이가 많아지면 염기서열 조립 시 염기서열의 연속성(continuity)이 떨어져 지게 되며, 특히 이배체 다염기변이 유전체의 경우는 하나가 아닌 두 개의 일배체형에 나온 서열임을 고려하지 않고 있으므로 정확성 및 연속성에 있어 문제가 되고 있다.

2.2 전통적 WGSS 방법

WGSS 방법는 이미 언급했듯이 유전체를 무작위로 자른 작은 조각들을 염기서열 판독 기술을 이용하여 서열화하고, 이 서열에서 서로 중첩되는 부분을 연결하면서 유전체를 조립하는 방법이다. 이는 DNA로부터 최대 500~800 bp 염기서열을 판독할 수 있는 Sanger's Method의 실험적 한계성으로 인해 이 최대 길이보다 큰 길이를 가진 유전체의 염기서열을 결정하기 위해서 고안되었는데, 알고리즘의 개요는 다음과 같다.

대략 ~3 Gb에 달하는 인간의 유전체와 같은 큰 유전체의 염기서열을 결정하기 위해서는 일단 염기서열을 결정할 유전체 7~8개(coverage)의 무작위 위치를 절단하여 수많은 2~3 kb 크기의 조각으로 만든다. 일단 2~3 kb 크기로 잘라진 조각들은 실험기법을 사용하여 양쪽 끝 500~800 bp정도는 염기서열을 결정할 수 있게 된다. 이 때 서열이 결정된 500~800 bp의 조각들을 read라고 부른다. 그 다음에 이런 염기서열이 결정된 작은 조각들을 서로 비교하여 유사한 조각들을 이어 붙여나가게 된다. 자세한 과정은 (그림 2)에서 볼수 있으며, 이런 과정을 서열 조립(sequence assembly)이라 부른다. 이런 방식으로 염기서열을 결정하는 프로젝트를 흔히 지놈프로젝트라고 부르는데, 대부분의 경우에 지놈프로젝트의 결과물은 유전체의 연속적인 전체 염기서열이 아니라, 불연속적인 scaffold들의 집합이다. 각각의 scaffold의 길이는 수십 Mb부터 수 kb까지 매우 다양할 수 있으며, 서로 다른 scaffold는 원칙적으로 전체 유전체상에 중첩되지 않고, 다른 scaffold사이에는 우리가 알 수 없는 작은 길이의 공간(gap)이 존재한다. 지놈프로젝트가 성공적이기 위해서는 밝혀진 염기서열도 정확해야 하지만 프로젝트를 통해 산출된 scaffold의 크기가 충분히 커야한다.

Scaffold의 크기가 충분히 클 경우에 우리는 scaffold가 연속성 측면에서 우수하다고 하고 성공적인 지놈프로젝트라고 평가할 수 있다. 가장 이상적인 경우는 하나의 scaffold가 하나의 염색체인 경우지만 이런 경우는 7~8개의 DNA를 사용하는 경우에는 불가능하다. 극단적으로 안 좋은 경우는 하나의 scaffold가 불과 2~3kb인 경우이지만 이 경우도 일어나기 힘든 경우이다. Scaffold들이 연속성측면에서 우수하지 않으면, 지놈프로젝트 자체가 의미가 없으며 많은 문제점이 생길 수 있다. 가장 대표적인 것은 scaffold길이가 짧아지면 유전자를 예측하기가 힘들어진다는 것이다.

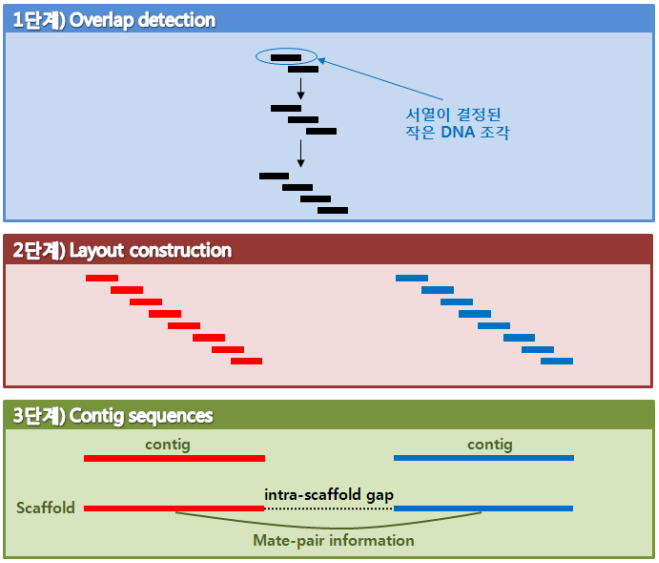

그림 2. 유전체서열 조립 (sequence assembly)은 대략 세 가지 단계로 나누어진다. 1단계에서는 서열이 결정된 작은 조각들 (read)을 서로 비교하여 유사한 조각들끼리 붙여 나간다. 이때 2개 이상의 read들이 합쳐진 결과물을 contig라 부른다. 2단계에서는 이어 붙인 contig들에 multiple alignment를 적용해서 assembly layout을 만든다. 3단계에서는 contig들로부터 consensus base를 도출하여 consensus sequence를 만들어낸다. 2~3 kb 크기의 각각의 DNA조각들의 염기서열을 결정할 때, 양끝의 500~800 bp의 염기서열을 동시에 결정하게 된다. 그러므로 두 개의 read들이 하나의 DNA조각에서 유래되었다는 정보가 추가로 주어질 수 있다. 이 정보를 mate-pair information이라고 한다. 만약 두 개의 분리된 contig들 사이에 mate-pair information이 존재할 경우에 두 contig들은 전체 DNA상에서 상당히 가까운 부분에 위치하리라 유추할 수 있으므로 두 개의 contig를 합치며 대략적인 공간을 두 개의 contig사이에 넣어준다. 이 합쳐진 contig들을 scaffold라 부른다.

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 유전체조립에 관련되어서는 현재 많은 진전이 있어왔고, 유전체 조립의 연속성 측면에서도 우수하다. 현재 인간 유전체의 N50 scaffold length는 2.7 Mb이고 침팬지 유전체의 경우에는 N50 scaffold length가 2.3 Mb정도 되는 것으로 알려져 있다[14]. 현재 공개된 다염기변이 유전체 서열들 중에는 시스템생물학의 연구에 있어서 아주 중요한 생물들인 멍게와 성게의 유전체 서열이 있지만 그런 생물들의 염기서열의 연속성(continuity)에는 심각한 문제점이 존재하고 있다.

3. 서열 정렬 툴

서열 정렬은 서열간의 상관관계를 보여주기 위해, 특히 상동성(homology)을 나타내기 위해 염기서열이나 단백질 서열을 정렬하는 것을 말하며, 관심대상인 하나의 서열과 상동성이 높은 서열들을 알아내어 서열의 기능을 유추하거나, 관련있는 서열들간의 진화적인 상관관계나 관련기능 부위 등을 예측하기 위한 목적으로 사용된다[7]. 기존에 알려진 서열을 대상으로 하는 상동성 검색(homology search)을 통해 연관성 있는 서열의 전부나 일부분을 알아내는 것은 생물학자들이 가장 빈번히 사용하는 분석으로, 이를 위해 dynamic programming을 이용한 pairwise alignment 알고리즘이 개발되어 사용되었다. 그러나 이는 데이터베이스를 대상으로 하는 상동성 검색에서는 실행시간이 $O(n^2)$으로 너무 오래 걸린다는 문제점을 갖고 있다. 또한 점차 local homoloy 검색에 대한 필요성이 대두되면서 heuristic을 이용한 빠른 알고리즘들이 개발되었다. 이렇게 하여 현재까지 개발되어 온 서열 정렬 툴들의 기본 아이디어는 WGSS 방법의 중첩 추출 단계와 layout contruction 단계에서 응용되고 있다. 본 논문에서는 이러한 서열 정렬 툴들 중에서 가장 많이 쓰이고 있는 FASTA[1][7], BLAST[2][7], BLAT[3]의 알고리즘에 대하여 분석해 보고자 한다.

3.1 FASTA

FASTA 알고리즘은 4가지 단계로 이루어지며 핵심은 해싱(hashing)을 사용하여 질의 서열과 상동성이 있는 부분을 O(n) time에 찾아내고 가장 상동성이 높을 것으로 예측되는 부분에 대해서만 정렬을 시행하는 것으로 알고리즘의 개요는 다음과 같다.

1) 첫 번째 단계에서는 질의 서열에 대해 lookup table을 구성하고 데이터베이스에서 각 비교 서열을 추출하여 k tuple이 일치하는 것을 찾고 이중 최대 일치를 보이는 10개의 diagonal region을 찾아낸다.

2) 두 번째 단계에서는 첫 단계에서 찾아진 diagonal region에 대해 scoring matrix를 이용하여 mismatch를 포함한 score를 다시 계산하고 maximal score를 가지는 subregion을 initial region으로 한다.

3) 세 번째 단계에서는 특정 threshold값 이상의 initial region만 대상으로 하여 그 위치와 각각의 score, joining penalty를 이용하여 initial region들을 연결하여 최적 정렬 서열(optimal alignment sequence)를 구한다.

4) 네 번째 단계에서는 세 번째 단계에서 찾아진 서열들에 대해 가장 높은 score를 가지는 initial region 주위에서 banded smith-waterman 알고리즘을 통해 optimal alignment를 구하게 된다.

3.2 BLAST

BLAST는 FASTA와 같이 heuristic을 이용하여 상동성을 검색하는 툴로 알고리즘의 개요는 다음과 같다.

1) 질의 서열을 분석하여 질의 sequence의 <W-mer>와 score T이상의 상동성을 가지는 <W-mer>를 모두 생성한다.

2) 데이터베이스의 각 비교 서열을 스캔하면서 이미 구성된 <W-mer>와 일치 되는 것이 있는가를 찾는다. 일치된 <W-mer>들 중 같은 diagonal에서 일정 distance이내에서 single hit들이 겹치지 않으면서 2개 존재할 때만 부분 정렬(local alignment)를 위한 seed 서열로 사용된다.

3) 이렇게 선택된 각 seed 서열에 대해 질의 서열과 비교 서열의 좌우를 확장(extension)해 간다. 확장해 가는 도중 최대치 값보다 일정값 이상 차이가 나는 경우에 확장을 중단하고 이를 HSP(high scoring pairs)로 한다.

4) 특정 cutoff값인 Sg 이상인 HSP에 대해 갭을 포함한 확장(gapped extension)을 수행하여 MSP(maximal segment pair) 구한다. 이 MSP에 대해 특정 cutoff 값 S 이상이면 이를 저장한다.

3.3 BLAT

BLAT은 BLAST와 비슷한 서열 정렬 툴로 제안된 것으로 다른 툴과의 가장 큰 차이는 질의 서열인 아닌 서열 데이터베이스를 중첩되지 않는 k-mer로 인덱싱(데이터베이스 상의 위치도 함께 저장한다.)하여 RAM에 올려놓고 질의 서열을 리니어하게 스캔하면서 오버랩되는 k-mer hits를 찾는 다는 점이다. 이는 데이터베이스 스캔 시간을 상당히 많이 단축시킬 수 있다. BLAT은 크게 두가지 stage로 구성되는데, 알고리즘의 개요는 다음과 같다.

1) search stage에서는 기본적으로 두개의 11-mer가 match되면 hits list에 포함시키고, 만들어진 hits list를 데이터베이스 상의 위치에 따라 64k bucket으로 분할하여 대각선에서 정렬(sort)시킨다. 이들 중 갭 한계(gap limit) 안에 있는 hits를 proto-clump로 묶는다. 이제 proto-clump내에 있는 hits를 데이터베이스 좌표(coordinate)에 의해 정렬(sort)하고 window limit 내에 있는 것들만 포함하여 real clump로 만든다. 여기서 최소 hits 수보다 적은 hits를 가지는 clump는 버리고 나머지 clump만을 선택한다.

2) alignment stage는 search stage에서 선택된 상동 지역(homologous regions)과 질의 서열과의 hits list를 생성하면서 시작된다. 이 과정에서 질의 k-mer와 상동지역에 여러개의 k-mer의 match가 존재하면, 매치가 하나가 되게 되거나 특정사이즈를 초과할 때까지 반복적으로 k-mer를 확장한다. 즉, mismatch가 없는 한 가장 길게 중첩되는 hits를 머지(merge)하는 것이다. 만약 여러개의 매치가 뒷받침된다면 하나 또는 두개의 mismatch를 허용하는 확장을 허용한다.

4. 다염기변이 유전체와 서열 정렬 툴

BLAST는 FASTA는 기본적으로 상동성 검색시간을 줄이기 위한 heuristic을 사용한 알고리즘으로 banded smith-waterman 알고리즘을 통해 optimal alignment를 구하는 방법이다. 일반적으로 FASTA가 BLAST에 비해 좀더 엄격한 smith-waterman 알고리즘을 수행하므로 좀더 좋은 정렬 결과를 내지만 계산량이 많게 되므로 수행시간이 오래 걸리게 된다.

BLAT이 BLAST나 FASTA에 비해 가장 큰 장점은 질의 서열이 아닌 데이터베이스를 인덱스화하여 RAM에 올려 사용한다는 점이다. 이는 서열 정렬 시간을 단축시키는데 큰 역할을 하지만, BLAT의 알고리즘은 90%이상의 염기 동질성(identity)를 가질 경우에만 좋은 결과를 줄 수 있다.

다염기변이 유전체의 경우, 멍게나 성게와 같이 실험적으로 DNA를 추출할 수 있는 유전체들도 있지만, 수많은 난배양 미생물(VBNC, Visible But Non-Culturable)의 경우에는 기존의 순수배양 환경에서는 배양이 되지 않는다. 이런 미생물들은 연구실 환경에서 유전체 조립에 필요한 샘플을 확보할 수 없기 때문에 특정 환경에 존재하는 생물 집합 전체의 유전체를 재취하여 유전체 서열을 조립해야 한다. 이 경우 다른 생명체에서 온 대용량 유전체들을 한꺼번에 처리해야 하므로 수행시간이 한 생명체에 대한 정렬보다 더욱 중요한 이슈가 된다.

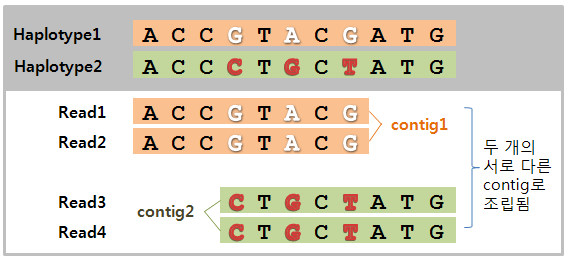

그림 3 다른 일배체형으로부터 유래한 read들이 서로 다른 contig들로 조립되는 경우.

또한 전통적인 WGSS 방법을 사용하게 되면, 중첩 추출(overlap detection) 과정에서 read들의 서열 정렬이 적합하지 않게 되어, 다른 일배체형에서 유래한 DNA조각들이 하나의 contig로 합쳐져서 조립되지 못하고 자주 다른 contig로 조립되게 되거나(그림 3), 서로 병합되지 않아야 하는 read들이 하나의 contig로 잘못 병합되는 경우가 많이 생긴다(그림 4). 이런 문제점을 해결하기 위해서는 일배체형에 대한 고려와 함께 identity level cutoff 조절이 필요하다.

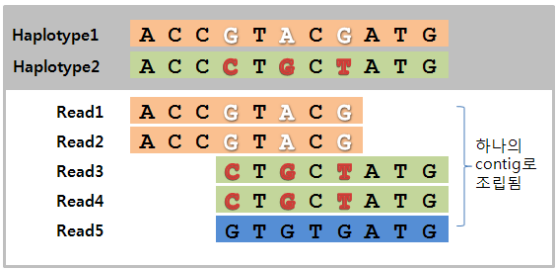

그림 4 서로 병합되지 말아야 하는 read들이 하나의 contig들로 병합되는 경우. Read 5는 다른 contig로 병합되어야 함.

그러므로 염기변이가 많은 다염기변이 유전체 조립을 위해서는 BLAST에 기반한 알고리즘을 기초로 하여 BLAT에서 제안된 데이터베이스 인덱스화 방법을 도입할 수 있다면 수행시간을 줄이는 것이 가능할 것이다. 또한 중첩 추출 과정에서 적합하지 않은 서열 정렬에 의한 잘못된 병합을 방지하기 위하여 identity level cutoff를 조금 낮추어 염기변이를 수용할 수 있게 해야하는 한편, 일배체형간 분리를 통해 염기서열의 연속성을 높일 수 있도록 일배체형 유추 방법도 도입해야 할 것이다.

5. 실 험

본 논문의 실험은 이미 살펴본 세 가지의 서열 정렬 툴인 BLAST, BLAT, FASTA의 수행시간이 다염기변이 유전체에 대해서 예상과 같은 결과를 보이는지 알아보기 위한 것이다. 즉, 다염기 변이 유전체인 멍게들(*Ciona intestinalis*, *Ciona savignyi*)과 성게(*Strongylocentrotus purpuratus*)의 유전체에서 각각 1000개의 read를 랜덤(Random)하게 뽑아서 각 서열 정렬 툴에 대해 정렬에 소요되는 시간을 초단위로 측정하였다. 실험 결과를 살펴보면 예측된 대로 BLAT이 가장 빠른 결과를 보였음을 알 수 있다.

(단위:초)

	성게 (Strongy locentrotus purpuratus)	멍게 (ciona savignyi)	멍게 (ciona intestinalis)
BLAST	1642	2850	1531
BLAT	1017	1662	1088
FASTA	76861	462908	19510

표 1 BLAST, BLAT, FASTA의 멍게들(*Ciona intestinalis*, *Ciona savignyi*)과 성게(*Strongylocentrotus purpuratus*) 유전체 read에 대한 서열 정렬 수행 시간 비교

6. 결 론

지금까지 다염기변이 유전체를 위한 유전체 서열 조립 전략 수립을 위해 이의 기본이 되는 다염기변이 유전체, 전통적 WGSS 방법, 서열 정렬 툴의 특성을 알아보

았다. 결국 다염기변이 유전체에 적합한 조립 알고리즘은 유사 조각을 찾아 연결하는데 걸리는 수행시간을 줄여야 하며, 중첩 추출 과정에서는 identity level cutoff를 조금 낮추면서 일배체형 유추 방법을 도입해야 한다.

7. 참고 문헌

[1] Pearson, W.R., and D.J. Lipman, Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 1988.

[2] Altschul, S.F., et. al., Gapped BLAST and PSI-BLAST:a new generation of protein database search programs. Nucleic Acids Res. 25, 1997.

[3] W. James Kent, BLAT-The BLAST-Like Alignment Tool. Genome Research, 2002

[4] The Sea Urchin Genome Sequencing Consortium, 2006 The Genome of the sea urchin *Strongylocentrotus purpuratus*. Science ,2006

[5]Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., et al. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. Science , 2002.

[6]Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., et al. Arachne: A whole-genome shotgun assembler. Genome Research , 2002.

[7]박기정,"서열 정렬 알고리즘과 유전체 연구에의 응용" 미생물학회지 제 34권 제1·2호, 1998.

[8]Kim, J. H., Waterman, M. S., Li, L. M. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. Genome Research, 2007.

[9]Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., et al. Community structure and metabolosim through reconstruction of microbial genomes from the environment. Nature, 2004.

[10]Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch D. B., et al. Environmental genome shotgun sequencing of the Sargasso sea. Science, 2004.

[11]Abdreas D. Baxevanis, B. F Francis Ouellette, Bioinfomatics : A practical guide to the analysis of genes and proteins Third edition 2005.

[12]Altschul, S.F. et. al., Basic local alignment search tool. J. Mol. Biol. *215*, 1990.

[13]Won, J., Park, S., Yoon, J., Kim, S. An efficient approach for sequence matching in large DNA databases, Journal of Information Science, 2006.

[14]Istrail, S., Sutton, G.G., Florea, L., Halpern, A., Mobarry, C.M., et al. Whole-genome shotgun assembly and comparison of human genome assemblies. Proc. Natl. Acad. Sci., 2004.