DNA Fragment Assembly

박 근수 (서울대학교 컴퓨터공학과 교수)
(02-880-8381, kpark@theory.snu.ac.kr)

Abstract:
최근 인간 지놈(genome)의 DNA가 밝혀져서 많은 관심을 받았는데, 이를 수행하는 방법을 소개한다. Human Genome Project에서 채택한 BAC-to-BAC 방식과 Celera 회사에서 채택한 whole genome shotgun 방식을 설명한다. 또한 두 방식에서 공히 fragment assembly 프로그램을 사용하는데, 이 프로그램의 개요를 설명한다.

약력

83 서울대 컴퓨터공학과 학사
85 서울대 컴퓨터공학과 석사
91 미국 Columbia 대학교 박사
91-93 영국 런던대 King's College 조교수
93-현재 서울대 컴퓨터공학부 부교수

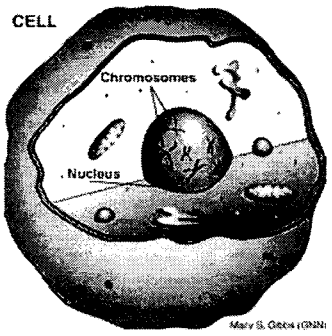
DNA Fragment Assembly

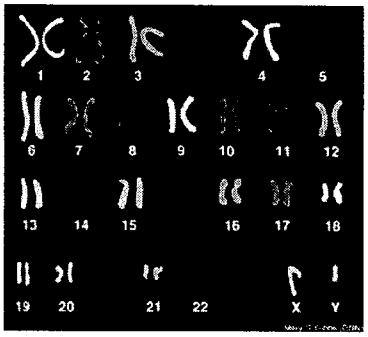
서울대학교 컴퓨터공학부
박 근 수

Overview

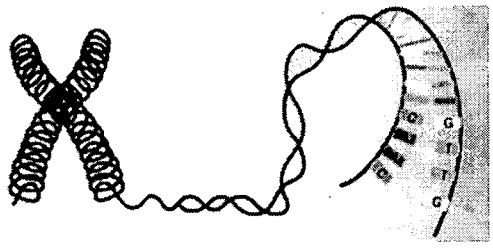
- Genome and DNA
- Sequencing Genome
- Assembling Genome

Chromosomes(염색체)



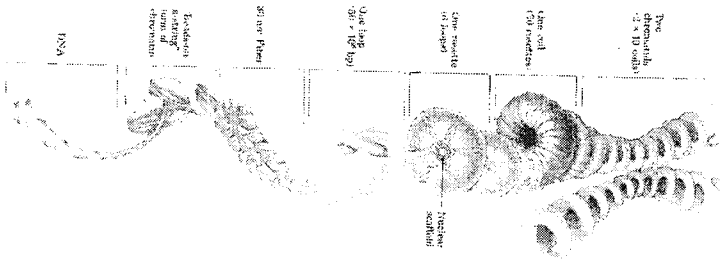


Chromosomes



- DNA in a human cell : 2m
- DNA in a human body : 2 X 10^{11} km
- Earth-to-Sun : 1.5 X 10^{8} km

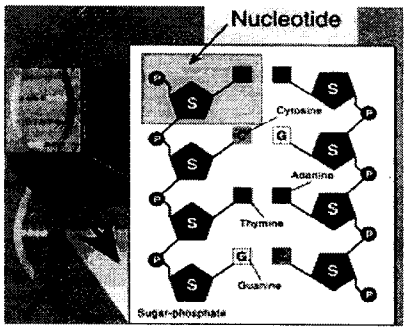
Chromosomes



- DNA : contents
- Protein : packaging

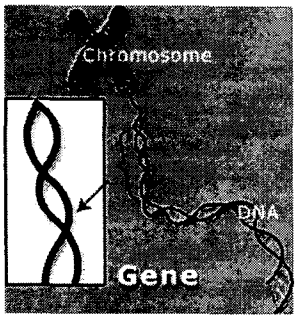
DNA(Deoxyribonucleic acid)

- Nucleotide들로 구성
  - Nucleotide = Sugar + Phosphate + Nitrogenous base
  - Adenine – Thymine
  - Guanine – Cytosine
- Double Helix 구조
- 인간의 DNA는
  약 $3.2*10^9$ base pairs
  로 구성



Gene and genome

- Gene(유전자)
  - Fundamental unit of heredity
  - 단백질을 합성하는데 필요한 정보 포함
  - Genome의 일부
- Genome(유전체)
  - 생명체가 갖는 전체 DNA



Overview

- Genome and DNA
- Sequencing Genome
- Assembling Genome

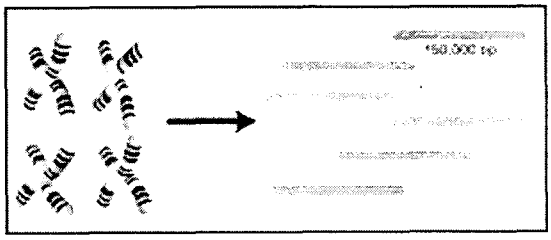
Genome sequencing

GTCCGCG T C GGCTCGGT

- BAC to BAC (Human Genome Project)
- Whole Genome Shotgun Sequencing (Celera)

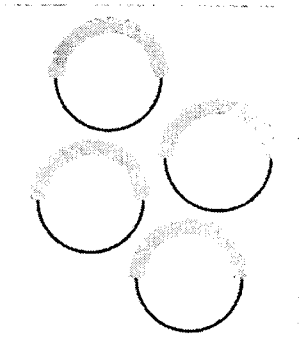
BAC to BAC

- Several copies of the genome are randomly cut into pieces that are about 150,000 bp long.



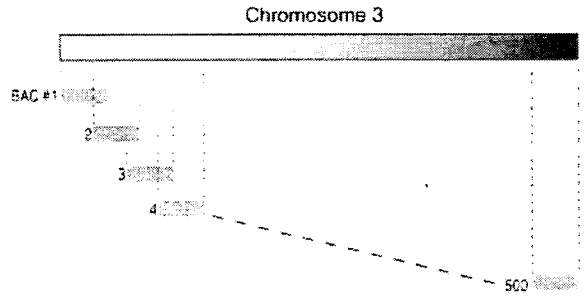
BAC to BAC

- Insert each of these 150,000 bp fragments into a BAC-a bacterial artificial chromosome.



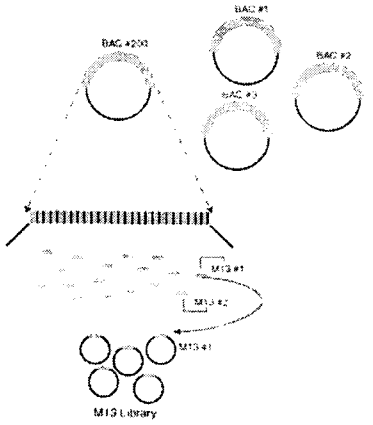
BAC to BAC

- Physical Map



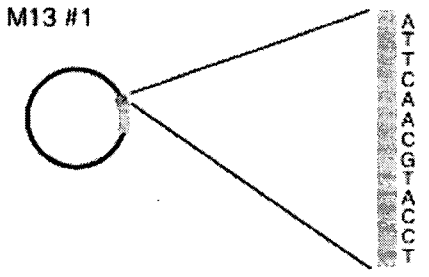
BAC to BAC

- Each BAC is broken randomly into 1,500 bp pieces and placed in M13.



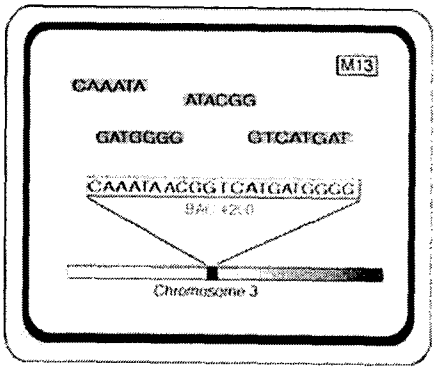
BAC to BAC

- 500bp from one end of the fragment are sequenced.

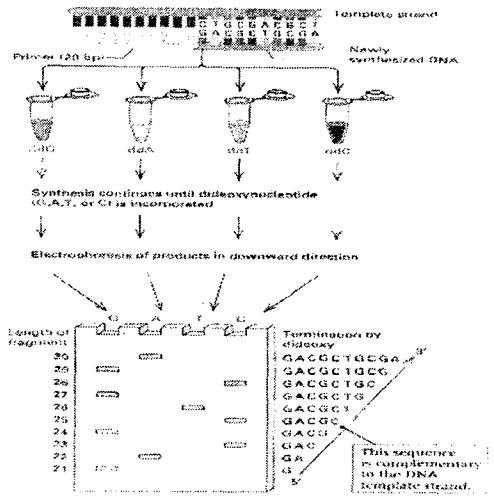


BAC to BAC

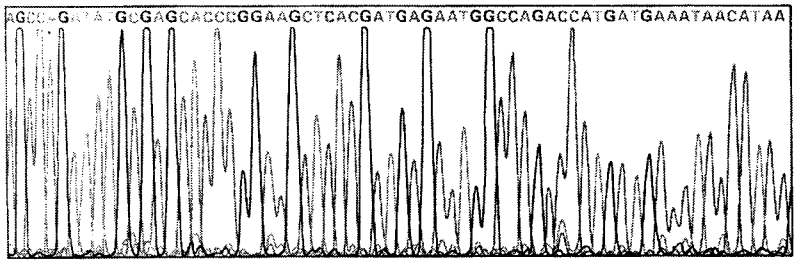
- Fragment Assembly (Phrap)



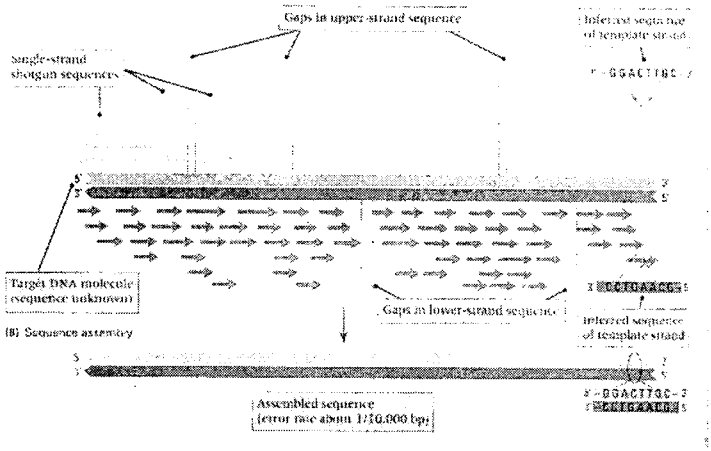
Gel Electrophoresis



Output from Sequencer

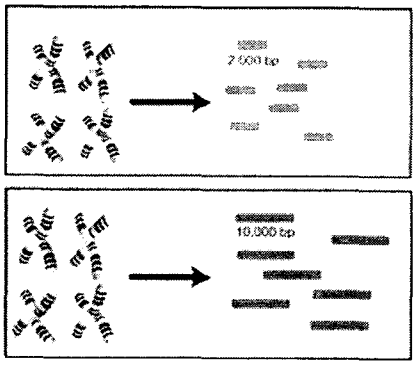


Shotgun Sequencing



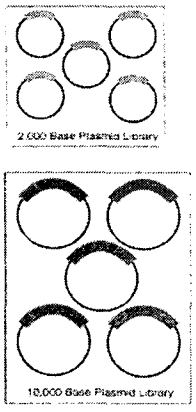
Whole Genome Shotgun

- Multiple copies of the genome are randomly cut into pieces that are about 2,000 bp and 10,000 bp long, respectively.



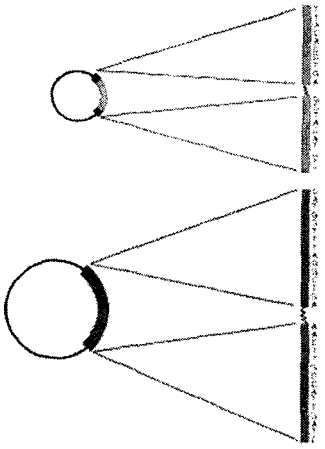
Whole Genome Shotgun

- Each 2,000 and 10,000 bp fragment is inserted into a plasmid.

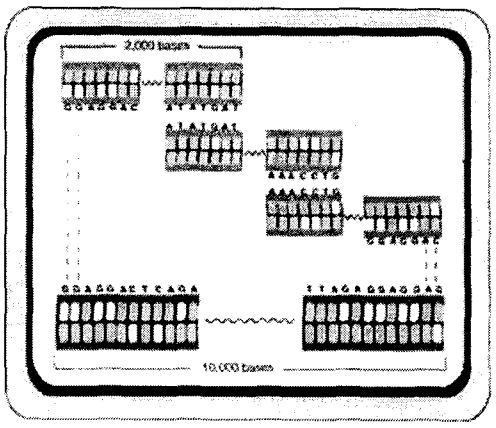


Whole Genome Shotgun

- 500 bp from each end of each fragment are decoded.



Whole Genome Shotgun



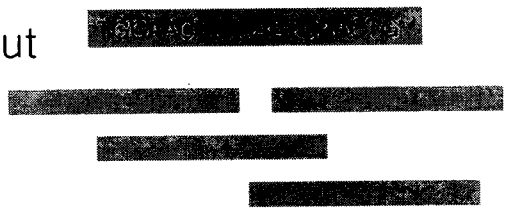
Overview

- Genome and DNA
- Sequencing Genome
- Assembling Genome

Fragment Assembly

- Pairwise alignment

- Layout



- Consensus

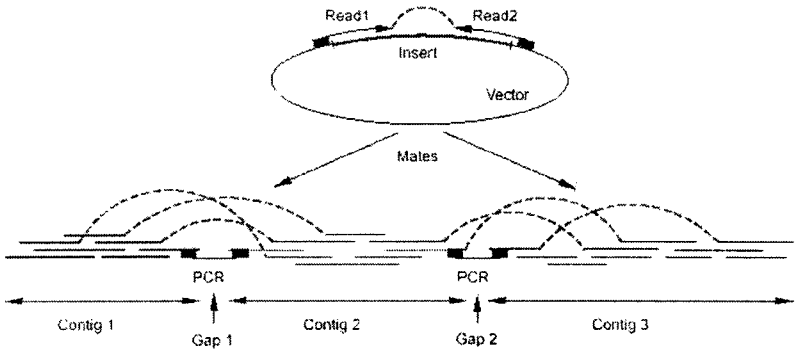
Pairwise Alignment

AATTGCATGCA

TGCAACTTTCACTGAACTG

ACTG ACTGTTTAC TTAC

Layout



Scaffold = {Contig 1, Contig 2, Contig 3}

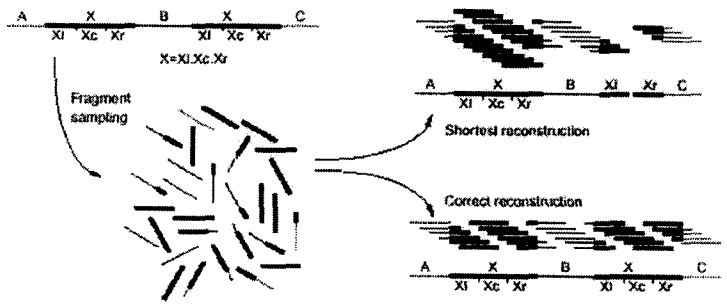
Consensus

A	*A*		*G*	***T***	*G*		***T***
A	*A*		–	***T***	*G*		***T***
–	*A*		*G*	*A*	*G*		***T***
A	*A*	–	*G*	***T***	*G*		–
A	*A*		*G*	***T***	*G*	*G*	***T***

Difficulties in Assembling Fragments

- Incomplete coverage
- Sequencing errors
- Unknown orientation
- Repetitive DNA

Repeat



Conclusion

- Development of DNA fragment assembly program that resolves these difficulties