박 현 석 박사

마크로젠 바이오인포매틱스 개발팀 이사 (& 세종대학교 조교수)

Tel. +82-2-3704-4611, Fax. +82-2-3704-4683

E-MAIL : hspark@macrogen.com, hsp@sejong.ac.kr

Address : 서울시 종로구 신문로1가 116번지 세안빌딩 9F <110-061>

◈ 연구관심분야

바이오인포매틱스, 인지과학

◈ 학 력

U. of Cambridge 전산학 박사, 1997

U. of Pennsylvanina 전산학 석사, 1994

서울대 전자공학과 학사, 1986

◈ 주요경력

(주) 마크로젠 이사, 1999

U. of Tokyo, 방문 교수, 1998

U. of Tokyo, Postdoctoral Fellow, 1997

U. of Pennsylvania, Research Fellow, 1993

◈ 연구 실적 요약

- 학술잡지 논문발표 : 5 편
- 학술컨퍼런스 논문발표 : 20 편
- 국제학회 기조연설, 초청강연, 초청세미나 : 20 회
- 저서/edited books : 1 권
- 연구과제 프로젝트 : 2 건

Whole Genome Fragment Assembly of Zymomonas mobilis and Its Metabolic Pathways

2002년 6월 27일
마크로젠
박현석

Introduction

- Genome Sequencing project
  - 생물체가 가진 유전체의 전체 염기 서열을 밝혀내는 작업.
  - 유전자들의 개수, 종류, 위치 등에 대한 제반 정보 확보.
  - 유전공학적인 활용을 위한 기초 자료 제공.
  - 병원성 미생물이나 산업적 가치가 풍부한 미생물들의 경우 genome project에 대한 중요성 증가.

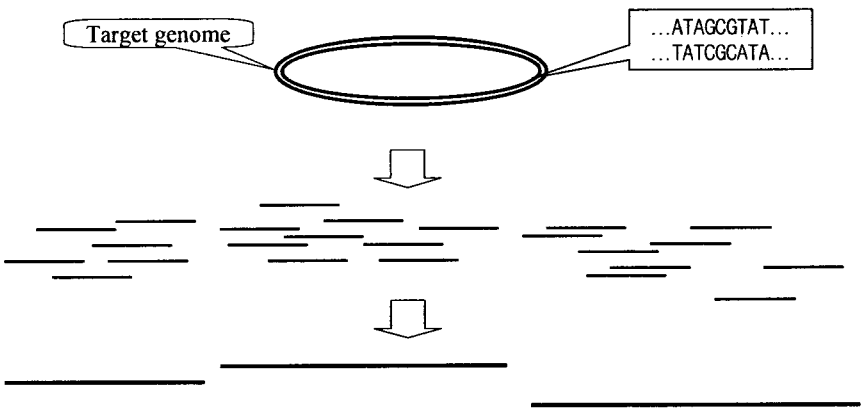
Introduction (*cont'd*)

- Random shotgun method
  - 1970년대 말에 Frederick Sanger 를 중심으로 한 연구진들에 의해 고안된 방법
  - 짧은 DNA조각을 조합하여 원래의 긴 연기 서열을 만들어 가는 방법
  - Pattern matching, dynamic programming, HMM model 등 다양한 전산학 알고리즘이 활용됨
  - 효모, 초파리, 인간, 쥐 및 70여 개 미생물들의 전체 염기 서열 분석

Whole-genome Sequencing Strategy

- Random small insert and large insert library construction
- Library plating
- High-throughput DNA sequencing
- Assembly
- Gap closing
- Editing
- Annotation

Random shotgun method



Random shotgun method (*cont'd*)

- 필요한 염기 서열 data의 길이
  - Shotgun 과정에서 만들어지는 DNA조각들은 모두 다른 DNA 조각에 의한 영향을 받지 않음.
  - DNA 조각의 양 끝의 염기 서열 data인 read들도 모두 서로에 대해 독립적이므로 각 read data 발생 확률은 포아송 분포를 따름.
  - 유전체 상의 임의의 염기 서열이 보유중인 read중에 존재하지 않을 확률은 $P_0 = e(-ln/G)$.

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

P_0: 분석되지 않은 염기일 확률.
l : read들의 평균 길이.
G : 전체 염기서열의 길이.
n: 클론 수.

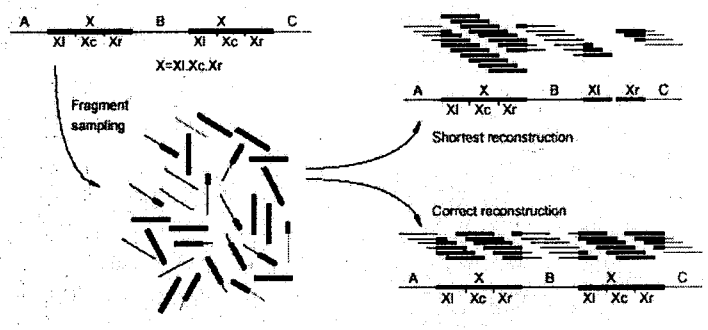
Difficulties

- Repeated/duplicated Region
- Lack of Coverage
- Sequencing Error
- Unknown Orientation
- Chimera

Difficulties

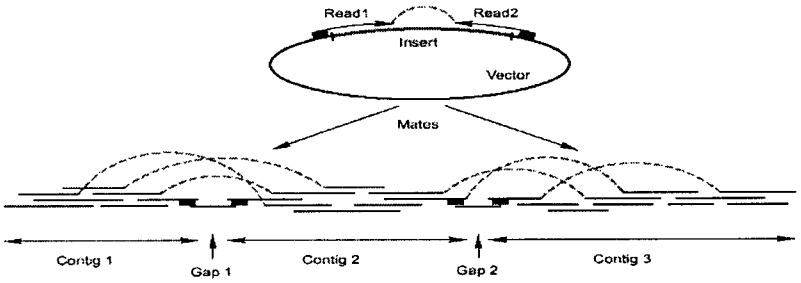
Repetitive / duplicated regions

- Repeated/duplicated Region



Difficulties
Using ***mate-pair*** information

- Mate-pair정보를 이용한 gap-filling 작업과 contig-ordering



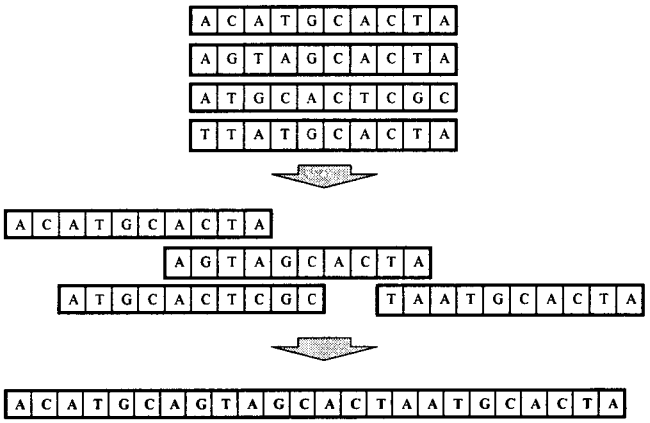
Fragment Assembly

- Popular assemblers
  - Phrap, TigrAssembler, CAP3, Celera Assembler
  - Usually implemented the 'overlap-layout-consensus' strategy
    1. Pair-wise alignment between every two sequencing-data(read).
    2. Laying out reads, merging reads overlapping another.
    3. Making consensus sequence from layout of reads.

Fragment Assembly

Assemble 과정

- Shortest Common Superstring problem -



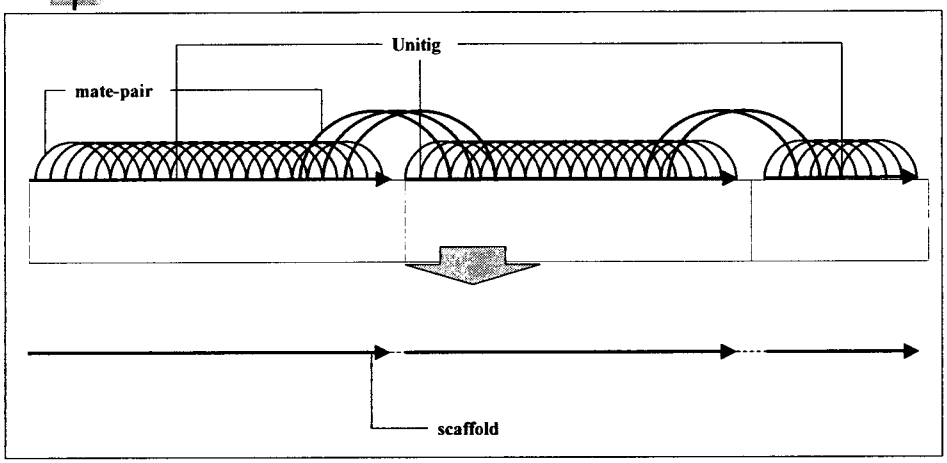
Fragment Assembly

주요 algorithm들 (*cont'd*)

- pair-wise comparison algorithm
  - Needleman-Wunsch algorithm (global alignment)
  - Smith-waterman algorithm (local alignment)
  - FASTA (William Pearson)
  - BLAST (altschul)
  - MSA(Multiple Sequence Algorithm)
  - PSI-BLAST
    - *In PHRAP*, banded smith-waterman algorithm.

Fragment Assembly

Unitig / Scaffold analysis (Contig ordering)



Fragment assembly of Z.mobilis

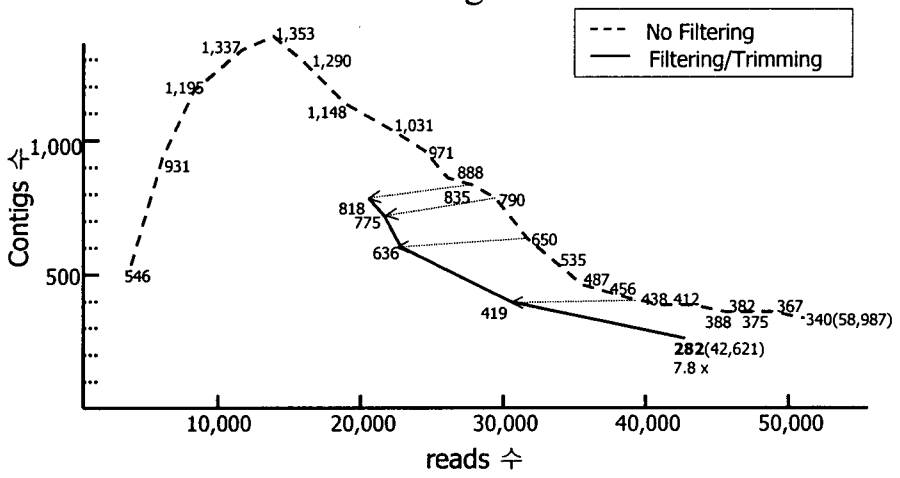
분석용 프로그램 제작

- Data 가공 및 분석 작업을 위한 프로그램들
  - 사용 언어 : Python / Perl
  - Data 가공 : 9개
    - Low-quality영역 제거, assemble상태가 양호한 부분 추출 작업 등.
  - Data 분석 : 20개
    - Read들의 배열 상태 파악, contig수 및 크기 분석, ordering 정보 추출, 비정상적인 read(chimera)들 구분 등.

Fragment assembly of Z.mobilis

1차 Assembly

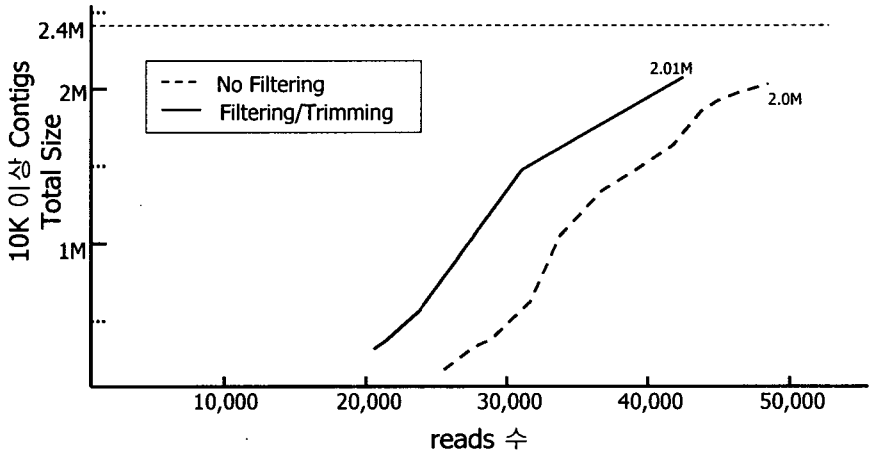
- read 수에 따른 contig 수 변화



Fragment assembly of Z.mobilis

1차 Assembly (*cont'd*)

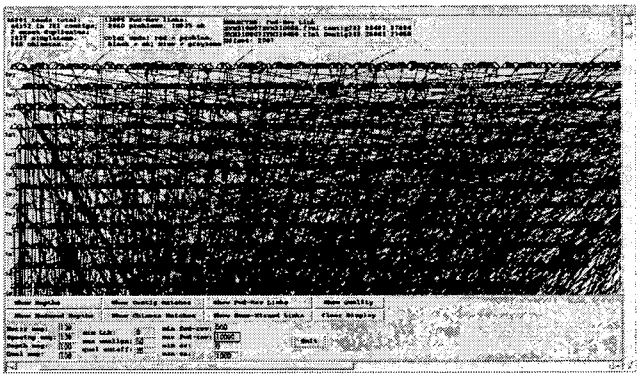
- 길이가 10K 이상인 contig들의 총 길이



Fragment assembly of Z.mobilis

1차 Assembly (*cont'd*)

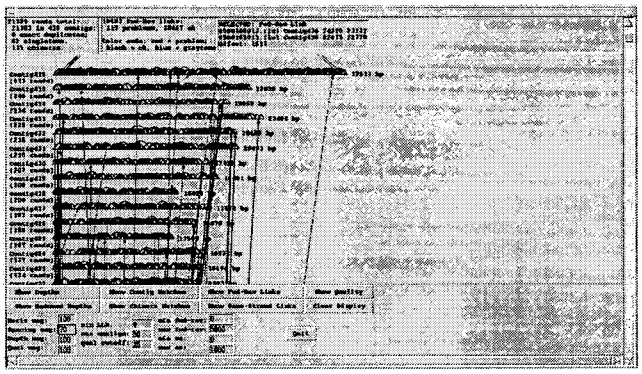
- 1차 assemble 작업 완료.
  - 282개 contig.
  - Mate-pair정보가 혼란스럽게 분포.



Fragment assembly of Z.mobilis

2차 Assembly

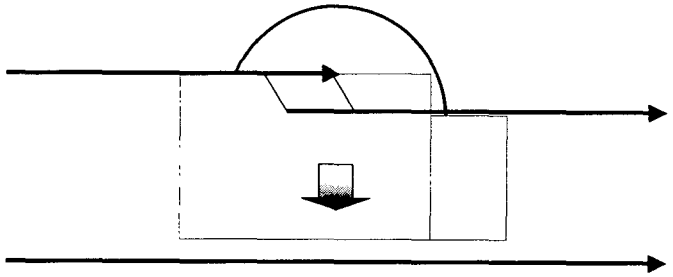
- Offset이 500bp~5000bp 이내에 있는 read들만
- 436개 contig.
- Mate-pair정보가 정돈됨.



Fragment assembly of Z.mobilis

최종 Contig

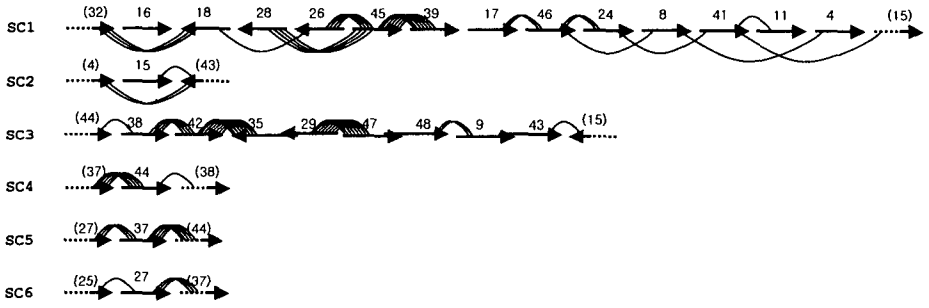
- Contig들 간의 유사도 정보와 mate-pair정보를 이용하여 8개의 임시 scaffold 제작.



Fragment assembly of Z.mobilis

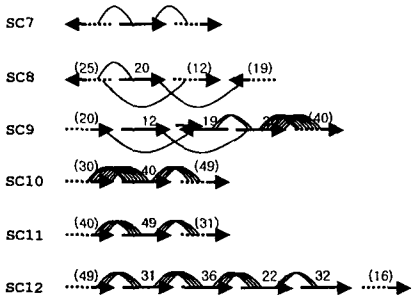
Contig - ordering

- 384쌍의 FOSMID data, *Pme*I, *Pac*I, *Not*I 및 23개 주요 유전자에 관한 physical-map사용.
- 12개의 ordered scaffold 생성.



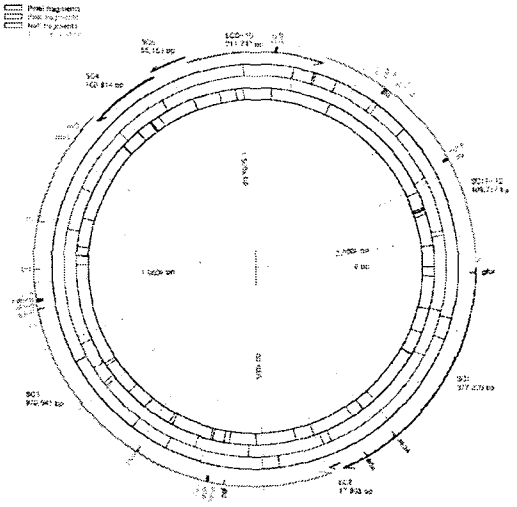
Fragment assembly of Z.mobilis

Contig - ordering (*cont'd*)



Fragment assembly of Z.mobilis

Contig - ordering (*cont'd*)



Fragment assembly of Z.mobilis

Annotation

- ORF prediction 소프트웨어를 이용하여 전체 genome sequence중 gene일 가능성이 높은 영역 추출.
  - Glimmer / GeneMark 등
  - HMM model을 사용함
- ORF sequence를 알려진 gene들의 sequence들과 비교하여 각 ORF들의 기능 추정.

Fragment assembly of Z.mobilis

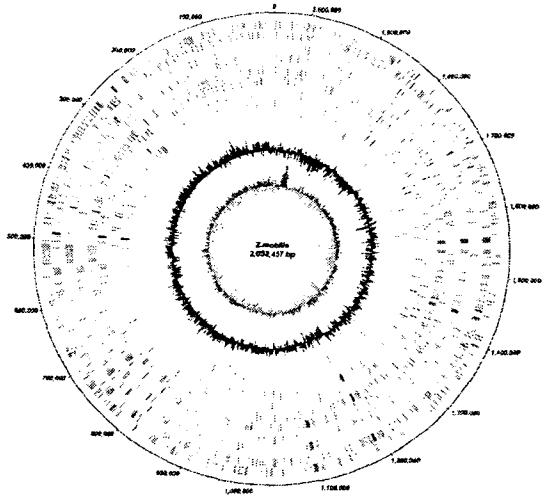
Annotation (*cont'd*)

- 2112개 유전자 발견

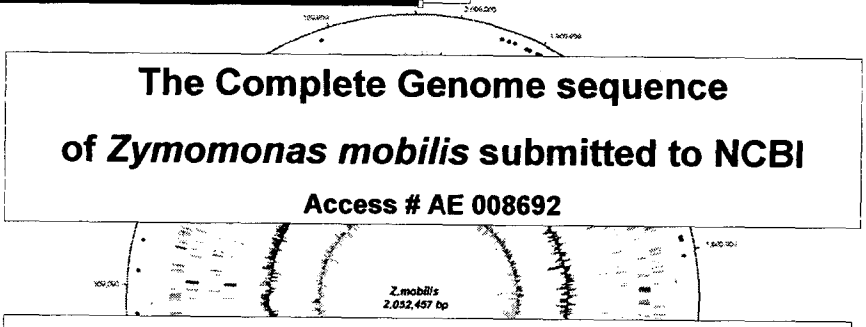
Information storage and processing	306
Cellular processes	412
Metabolism	531
Poorly characterized	863
total	2,112

Fragment assembly of Z.mobilis

Annotation (*cont'd*)



Microbial Genomics



The Complete Genome sequence

of *Zymomonas mobilis* submitted to NCBI

Access # AE 008692

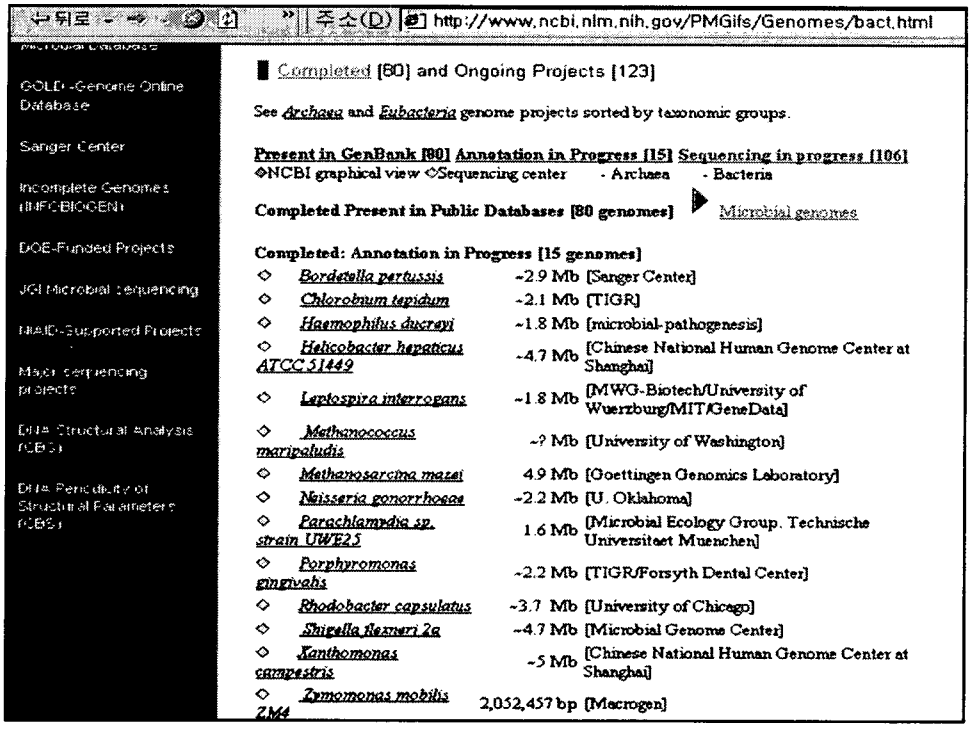
***Z. mobilis*: a valuable alcohol-producing microorganism**

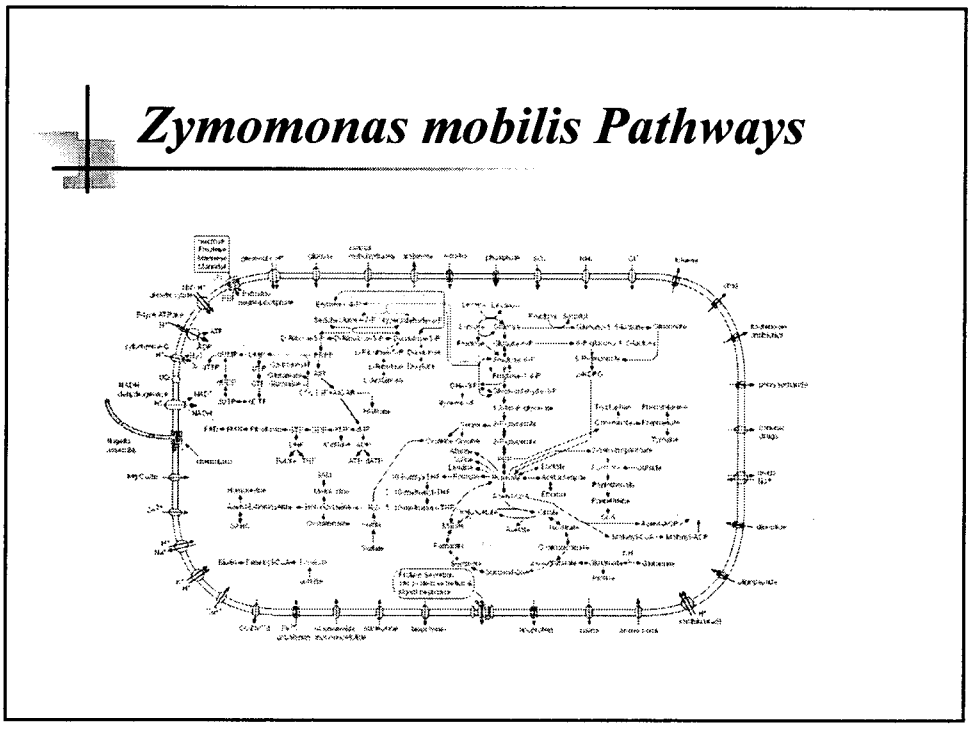
An alternative energy source for replacing costly fossil fuel

◎ Sequencing/Analysis of 2.1 milion base pairs of DNA

-Identification of 2,068 genes (34% unknown genes)

◎ Complete genomic analysis, developing DNA chips, processing patents


http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html
GOLD -Genome Online Database
Sanger Center
Incomplete Genomes (INFOBIOGEN)
DOE-Funded Projects
JGI Microbial Sequencing
NIAID-Supported Projects
Major sequencing projects
DNA Structural Analysis (CBS)
DNA Periodicity of Structural Parameters (CBS)
Completed [80] and Ongoing Projects [123]
See Archaea and Eubacteria genome projects sorted by taxonomic groups.
Present in GenBank [80] Annotation in Progress [15] Sequencing in progress [106]
◈NCBI graphical view ◇Sequencing center - Archaea - Bacteria
Completed Present in Public Databases [80 genomes]
Microbial genomes
Completed: Annotation in Progress [15 genomes]
◇ Bordetella pertussis ~2.9 Mb [Sanger Center]
◇ Chlorobium tepidum ~2.1 Mb [TIGR]
◇ Haemophilus ducreyi ~1.8 Mb [microbial-pathogenesis]
◇ Helicobacter hepaticus ATCC 51449 ~4.7 Mb [Chinese National Human Genome Center at Shanghai]
◇ Leptospira interrogans ~1.8 Mb [MWG-Biotech/University of Wuerzburg/MIT/GeneData]
◇ Methanococcus maripaludis ~? Mb [University of Washington]
◇ Methanosarcina mazei 4.9 Mb [Goettingen Genomics Laboratory]
◇ Neisseria gonorrhoeae ~2.2 Mb [U. Oklahoma]
◇ Parachlamydia sp. strain UWE25 1.6 Mb [Microbial Ecology Group. Technische Universitaet Muenchen]
◇ Porphyromonas gingivalis ~2.2 Mb [TIGR/Forsyth Dental Center]
◇ Rhodobacter capsulatus ~3.7 Mb [University of Chicago]
◇ Shigella flexneri 2a ~4.7 Mb [Microbial Genome Center]
◇ Xanthomonas campestris ~5 Mb [Chinese National Human Genome Center at Shanghai]
◇ Zymomonas mobilis ZM4 2,052,457 bp [Macrogen]


Zymomonas mobilis Pathways

Conclusions

- 의의
  - 한국 최초의 genome project.
- 응용분야
  - 대체 에너지 산업에 *Zymomonas mobilis*를 효율적으로 사용하기 위한 기초 자료.
- 효율적으로 read data를 assemble할 수 있는 알고리즘에 대한 지속적인 연구 필요.