정해영 박사
(주) 제노텍, 기술연구소 생물정보실 실장
Tel. 82-42-862-8404, Fax. 82-42-862-8406
e-mail : hyjeong@genotech.co.kr
address : 대전광역시 유성구 전민동 461-6 대덕바이오커뮤니티 <우:305-390>

◈ 연구관심분야 :
Microbial whole genome sequencing and analysis
DNA Chip

◈ 학 력
1. 한국과학기술원 생물공학과 학사, 1991
2. 한국과학기술원 생물과학과 석사, 1993
3. 한국과학기술원 생물과학과 박사, 1997

◈ 주요경력
2002 - 현재 (주)제노텍 기술연구소 생물정보실
1997 - 2000 한국과학기술원 생물과학과 Post doc.

◈ 연구 실적 요약
- 학술잡지 논문발표 : 1 편
- 학술컨퍼런스 논문발표 : 3 편
- 국제학회 기조연설, 초청강연, 초청세미나 : 1 회
- 특허 : 국내 1 건
- 연구과제 프로젝트 : 5 건

Strategies for Microbial Genome Sequencing, Assembly, and Annotation

정해영
㈜제노텍 생물정보연구실

최근 미생물 게놈 해석의 동향(I)

- GOLD (Genomes OnLine Database)
  - 16 a rchaeal genomes
  - 64 b acterial genomes
  - 2 83 prokaryotic ongoing genomes
- Integrated Genomics
  - 404 genomes integrated at ERGO (May 2002)
  - Only 6 genomes publicly available
  - Over 95% coverage, not necessarily one contig

GenoTech Corporation

최근 미생물 게놈 해석의 동향(II)

- Large genome size (>5 Mb)
- 방법론의 표준화
  - Modified shotgun strategy
  - Minimum two sets of libraries
  - Mate 정보를 이용한 contig 정렬 및 gap filling
  - Annotation methods

GenoTech Corporation

새로운 방법들

- Assembler program (주로 진핵생물용)
  - Euler : *Proc. Natl. Acad. Sci.* USA 98:9748 (2001)
  - Arachne : *Genome Res.* 12:177 (2002)
  - RePS : *Science* 296:79 (2002)
- Finishing methods
  - eOST, ASIN 등
  - Transposon mediated sequencing
- Automatic annotation

GenoTech Corporation

Modified Shotgun Sequencing

- 7~10X sequence coverage by end sequencing of short insert (typically 2~5 kb plasmid) library
- ~10X clone coverage by long insert library
  - Lambda, cosmid, fosmid, or BAC
- Utilization of mate information
  - For assembly verification and finishing

GenoTech Corporation

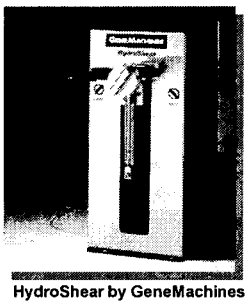
유전체 해석의 단계 (I)

- 라이브러리의 제작
- 클론 말단의 염기서열 결정
  - Large scale DNA preparation
  - Sequencing reaction optimization
  - Quality control
- Contig assembly

GenoTech Corporation

Sequencing Library의 제작

- Random fragmentation
  - Sonication
  - Nebulization
  - Hydrodynamic shearing
- Size selection by gel electrophoresis
- End-repair and ligation



HydroShear by GeneMachines

GenoTech Corporation

Library 제작 효율의 증대

- 평가 기준
  - Colony ti ter
  - White/blue c olony ratio
  - Clone fidelity
  - Clone stab ility
  - DNA prep efficiency
  - Randomness
  - Size dist ribution
  - Validity of m ate pair

➡ 시퀀싱 성공률을 좌우하는 가장 중요한 요인

GenoTech Corporation

배양 및 DNA prep

- 96-well type의 deep-well plate가 일반적
- 효율적인 aeration 및 well-to-well contamination의 방지가 중요
- Liquid handling을 위한 자동화 장비가 필수
- Alkaline miniprep 방법에 glass fiber-based multiwell filter plate를 통한 정제

GenoTech Corporation

DNA Sequencing

- Cycle sequencing using fluorescent dyes
- High throughput capillary electrophoresis
  - ABI Prism 3700 DNA Analyzer
  - MegaBACE
  - RISA
- 일반적인 반응 성공률은 85% 정도로 보고됨

GenoTech Corporation

Basecalling with Quality Value - PHRED

- 대용량 시퀀싱에서 가장 널리 쓰이는 basecalling software (*Genome Res.* 8:175, 1998)
  - cf. TraceTuner by Paracel
- 할당된 염기마다 0~99의 "Quality value"를 부여함

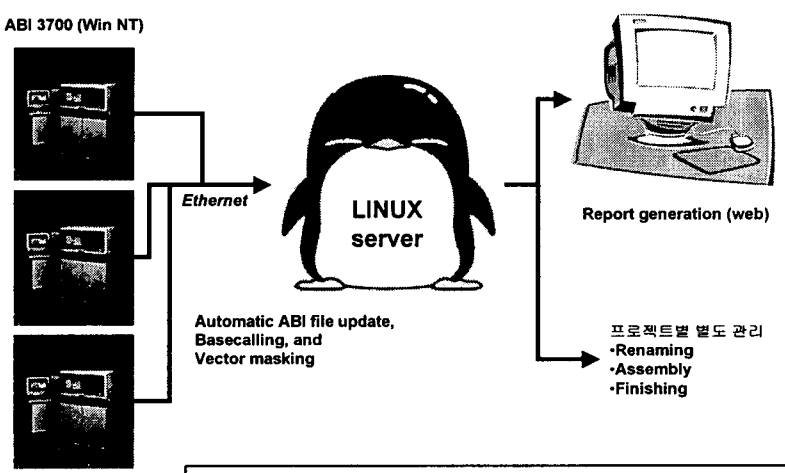
$$QV = -10 * \log P_e$$

P_e: error probability

- QV = 20: 정확도 99% (10^{-2} error)
- QV = 30: 정확도 99.9% (10^{-3} error)

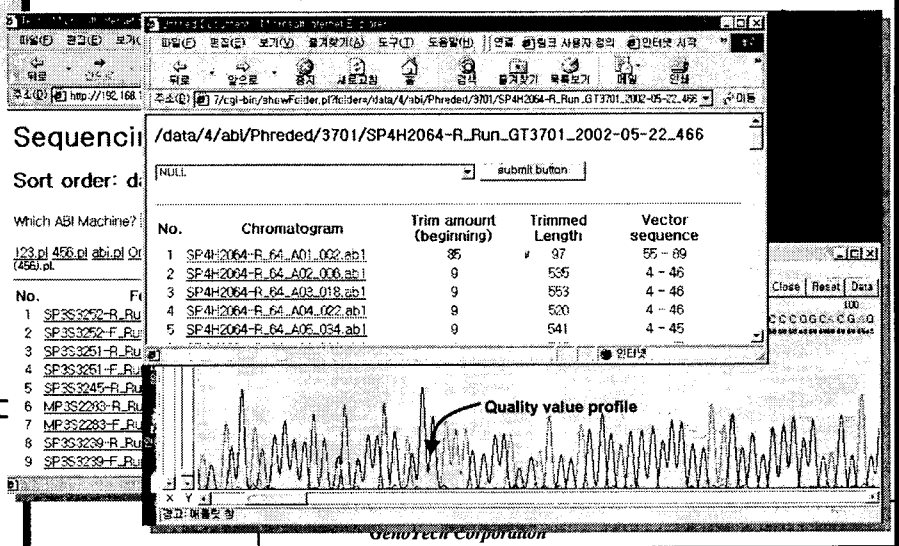
GenoTech Corporation

Data Management in GenoTech



GenoTech Corporation

시퀀싱 결과의 모니터링



Chromatogram의 처리 과정

- ABI file
  ↓ *Perl script*
- Renamed ABI file
  ↓ *Phred (basecaller)*
- PHD file
  ↓ *Phd2fasta*
- Fasta file and quality file
  ↓ *Cross_match against UniVec database*
- Vector-screend fasta file and quality file
  ↓ *Phrap (assembler)*
- ACE file
  ↓ *Consed (contig viewing and editing)*

University of Washington의 Software package 사용
http://www.phrap.org

GenoTech Corporation

Read Name Conversion

- St. Louis 규칙을 따름
- Read type, template (library) name, read direction 등의 정보를 담고 있어야 함
- Shotgun universal primer read
  - VV2S3123G01.b2.ab1
- Primer walking read
  - VV2S3123G01_321.020627.ab1

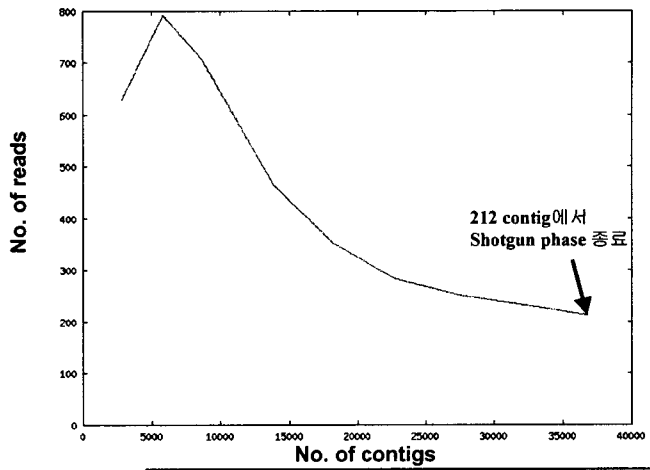
GenoTech Corporation

Shotgun Phase 중의 평가 항목

- 라이브러리 batch별 품질
- High quality base의 수
- Contig의 수 (급격히 증가후 감소)
- Contig 길이의 합 (유전체 크기에 수렴)
- Contig의 read 수 분포
- Singlet read와 short contig의 확인 (이종 세균의 오염 여부 확인)

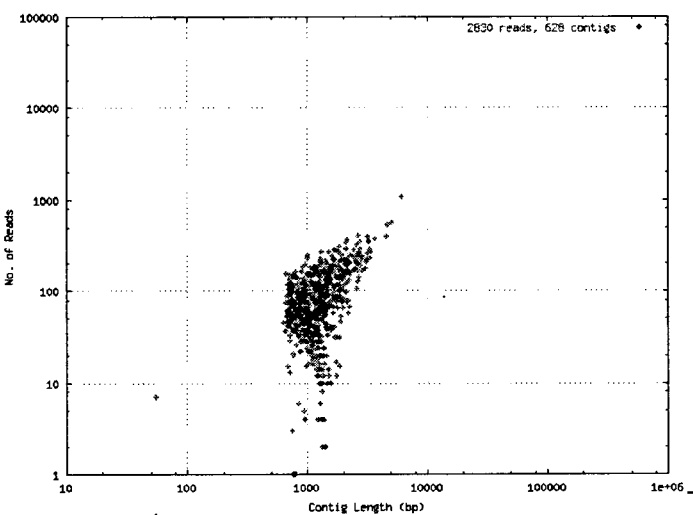
GenoTech Corporation

*Mannheimia sp.*의 Contig 수 변화



GenoTech Corporation

*Mannheimia sp.*의 Contig 크기 분포



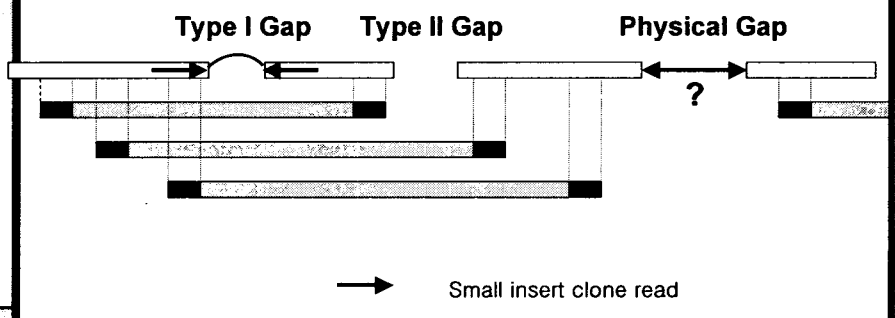
GenoTech Corporation

What is Mate?

- 한 subclone의 양 말단에서 얻어진 서열 정보의 쌍을 일컬음
- Contig의 순서를 결정하고 gap size를 추정하는 데 도움이 됨
- 실제 finishing 단계에서 walking용 template로 쓰임
- Contig 내에서 mate 관계에 있는 read는 간격이 clone size와 같고 서로 마주하고 있어야 함
- Assembly verification에 널리 쓰임

GenoTech Corporation

Contig, Scaffold (Supercontig) and Layout



GenoTech Corporation

유전체 해석의 단계 (II) - Finishing

- Gap이 없는 high-quality consensus를 얻어내는 작업
  - Gap filling
  - Assembly의 오류 확인
  - Consensus의 정확도를 일정 수준까지 개선
- 자동화하기 어렵고 수작업에 많이 의존
- 종료 시점을 예측하기 어려움

GenoTech Corporation

Finishing의 목표

- Coverage: no gap
- Accuracy: 99.99% (less than 1 error in 10 kb)
- Depth: at least two subclones for all nucleotide positions (PCR product 또는 genomic DNA에 대한 direct sequencing인 경우는 제외)

GenoTech Corporation

Gap Filling 방법

- Contig간을 연결하는 복수의 (sub)clone 탐색
- Contig 말단에 존재하면서 reverse read가 없는 경우 재반응을 유도할 수도 있음
- Library의 clone size가 다양할 경우 이를 반영하여야 함
  - Consed –autofinish (D. Gordon, Univ. of Washington)
  - Mapper (M. Zoddy, Whitehead Institute)
- Primer walking / mini-library / nested deletion / transposon-mediated sequencing

Calling missing reverse read at the end of contig (de novo)
GenoTech Corporation

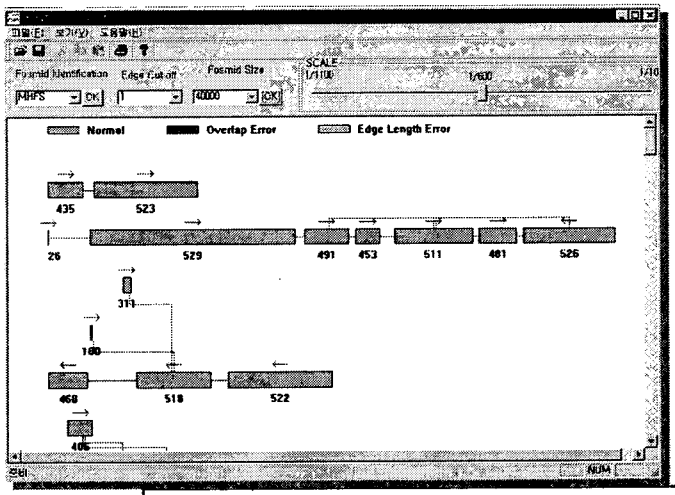
Autofinish의 활용 사례

Physical Gap의 해결 방안

- Combinatorial PCR
  - *Genomics* 62:500 (1999)
- Bioinformatic tools

GenoTech Corporation

ConPath (Scaffold Viewer)



GenoTech Corporation

GenoTech/SmallSoft

Finishing 작업의 어려운 점

- 적절한 finishing 개시 시점의 결정이 어려움. 특히 예산이 제한된 경우
- Genome의 내재적인 문제
  - Repeat sequence
  - Special chemistry를 요하는 경우 (high G+C 등)
- Contig의 수가 너무 많은 경우 다루어야 할 클론과 custom primer가 많아짐

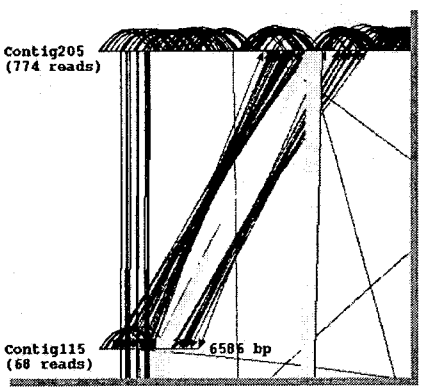
GenoTech Corporation

새로운 Gap-Filling Technology

- 제한된 primer library를 이용한 시퀀싱
  - eOST (Octamer-primed Sequencing Technology) : *Nucleic Acids Res.* 28:E22 (2000)
  - ASIN™ (Amplification and Ssequencing by Interlaced Nesting) : Takara Shuzo Co. (Japan)
  - UniSeq™ : Nucleics Pty Ltd. (Australia)
  - APA (Asymmetrical PCR Amplification) : Bio S&T Co. (Canada)

GenoTech Corporation

Assembly 오류의 예



- Phrap은 CAP3와 달리 assembly 과정에 mate 에 의한 constraint를 활용하지 않음
- Contig205는 invalid link 의 다발을 중심으로 둘로 나뉘어야만 함
- Consed에서 cut & rejoin 후 "fake read"를 추가하여 reassemble시 수정한 구조를 유지하게 함

GenoTech Corporation

Assembly Verification

- Scaffold analysis (mate information)
- Restriction fragments analysis
- 임의의 large insert clone을 골라 전체 서열을 결정한 뒤 genome 서열과 대조

GenoTech Corporation

이론과 실제의 차이는 어디에서 오는가?

- Library가 genome 전체를 반영하지 못함
  - Sampling 수준이 낮음 (Lander & Waterman prediction)
  - 라이브러리의 불균등 분포 문제
    - Fragmentation 방법의 한계
    - 재조합 클론의 불안정성(특히 미생물 genome)
  - Invalid mates가 많은 경우
- Sequencing quality
- Genome topology를 모르는 경우
  - Multiple (linear) chromosome, genome size 등
- Repeat sequence

GenoTech Corporation

유전체 해석의 단계 (III)

Annotation

The prediction of genes in a genome,including the location of protein-encoding genes, the sequence of the encoded genes, any significant matches to other proteins of known functions, and the location of RNA-encoding genes.

In Bioinformatics, David W. Mount (2001)

GenoTech Corporation

시퀀싱 오류의 정정

- Frameshift를 유발하는 coding sequence 내의 오류 정정
  - 상동성 비교에 의한 방법
  - Coding region의 본질적 특성에 의한 방법
    - Seqerr : http://www.imbb.forth.gr/seqerr.html
    - FSED : *Nucleic Acids Res.* 23:2900 (1995)
    - ProFED : *Genome Res.* 9:1116 (1999)

GenoTech Corporation

Gene Prediction 도구의 발전

- Intrinsic method: use statistics or pattern recognition algorithm to find genes through detection of specific motifs or global statistical patterns
  - GeneMark, Glimmer...
- Extrinsic method: use information derived from similarity search procedures
  - BLASTX, ORPHEUS...

GenoTech Corporation

Glimmer 2.02

- *build-imm* : 입력 서열로부터 IMM(Interpolated Markov Model)을 작성
- *glimmer* : IMM을 이용하여 전체 서열로부터 추정되는 유전자를 찾아냄. Overlapping gene의 자동 해결 기능 추가
- *RBSfinder* : 개시 코돈의 위치를 정확히 잡기 위해 ribosomal binding site를 검색

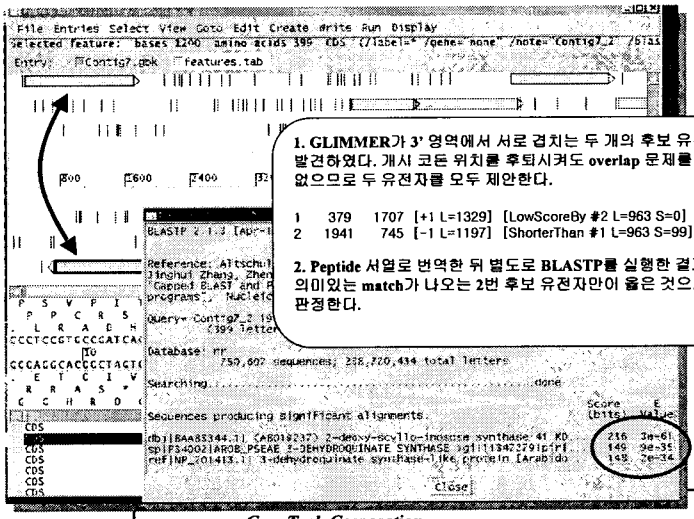
GenoTech Corporation

Annotation의 기본 전략

- Gene prediction
- 알려진 단백질에 대한 상동성 검색 (blast)
- COG category에 따른 기능 분류 (cognitor)
- Metabolic pathway 상에 할당 (KEGG)
- Protein hidden Markov Model에 대한 검색 (hmmer)
- RNA-encoding gene의 검색 (blastn, tRNAscan)
- Origin of replication 검색 (GC skew 이용)
- 비교유전체학적 분석
- DB integration, browsing system의 구성

GenoTech Corporation

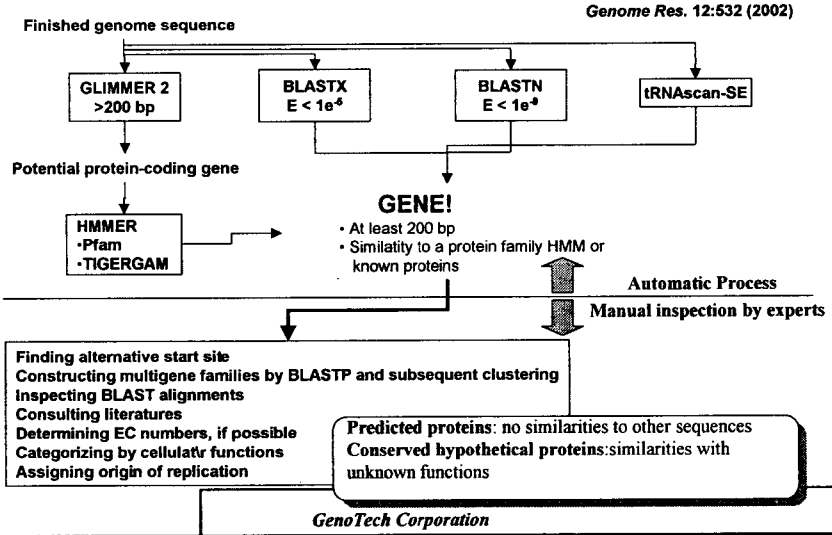
Artemis를 이용한 Annotation 사례



GenoTech Corporation

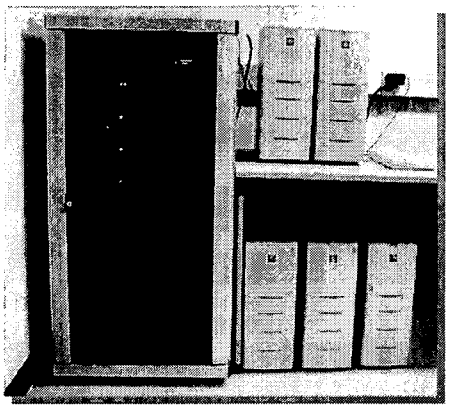
Calhoun Annotation System

Genome Res. 12:532 (2002)



GenoTech Corporation

Linux Cluster를 이용한 BLAST 분석

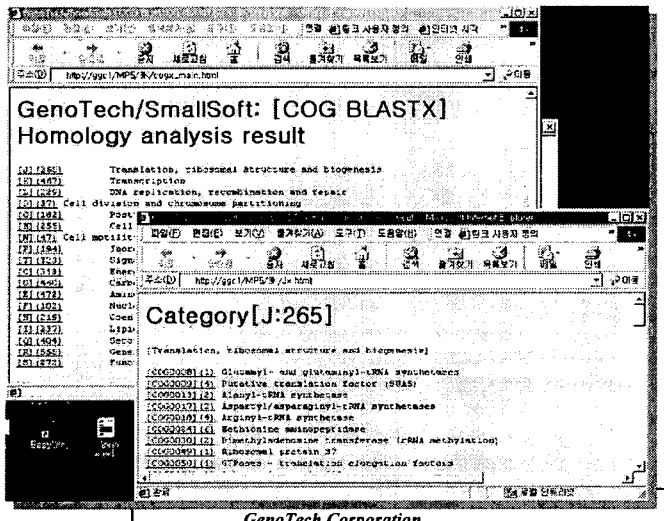


GGC Linux Cluster

- 5 P-III (1 GHz)
- 9 Alpha EV67 (600 MHz)
- Red hat 7.x
- Kernel 2.4.x

GenoTech Corporation

Functional Category Assignment (COG)



GenoTech Corporation

TIGR Gene Naming Rules (I)

- Information used for assignments
  - Pairwise search results
  - HMM matches
  - Prosite motifs
  - Active sites
  - Operon structure

GenoTech Corporation

TIGR Gene Naming Rules (II)

- High confidence

  enolase (eno)
- Suggestive, but not definitive evidences

  5-carboxymethy-2-hydroxymuconate delta isomerase, putative
- Specific assignment not possible, but member of a family

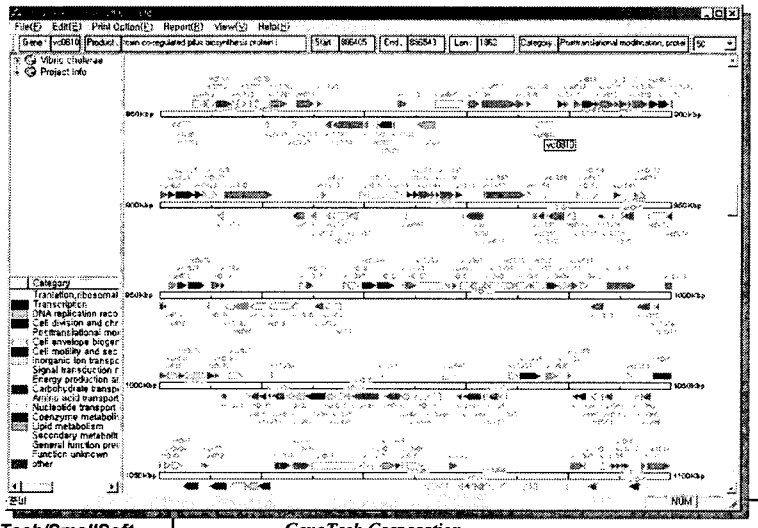
  carbohydrate kinase, FGGY family
- Similarity to the conceptual translation of gene

  HesB/YadR/YfhF family protein
- No database match

  hypothetical protein

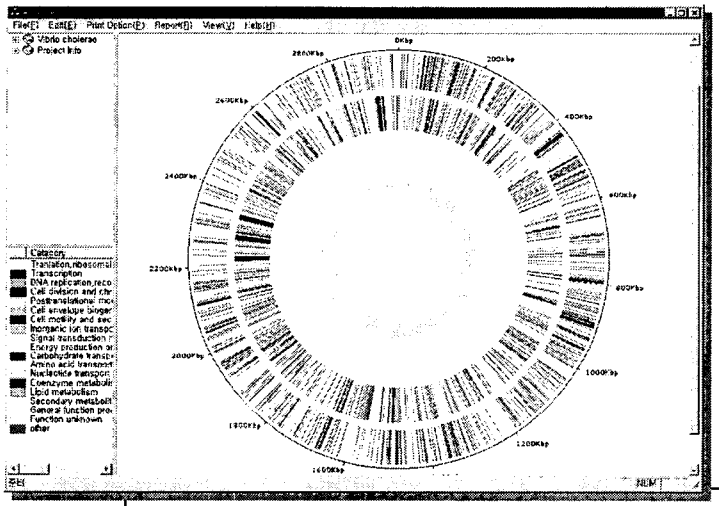
GenoTech Corporation

AnnoView (I)



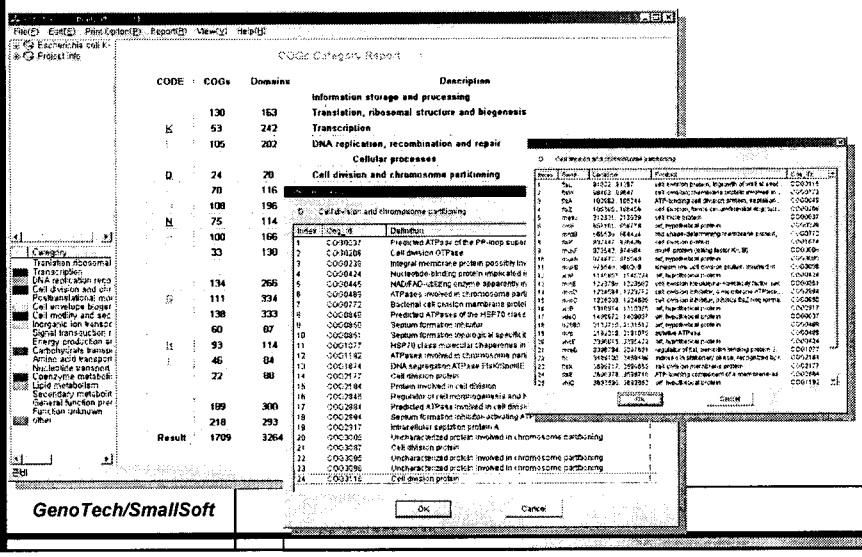
GenoTech/SmallSoft GenoTech Corporation

AnnoView (II)



GenoTech/SmallSoft GenoTech Corporation

AnnoView (III)



기타 Annotation 관련 도구들

- Pedant-Pro Sequence Analysis Suite (Biomax Informatics)
- Genotator (*Genome Res.* 7:754, 1997)
- GAIA (*Genome Res.* 8:234, 1998)
- AMIGA (*BMC Bioinformatics* 3:1, 2002)
- iANT (http://sequence.toulouse.inra.fr/R.solanacearum.html)
- genomeSCOUT (Lion Bioscience)
- airBASE! (Bioinfomatix, Inc.)

GenoTech Corporation

결론

- 국내의 미생물 유전체 해석 기술도 이제 경쟁력을 갖추어 나가고 있음
- 경제적인 finishing을 위한 새로운 기법 개발이 절실함
- Annotation 과정을 최대로 자동화 하되 전문가 그룹에 의한 검토가 반드시 필요함

GenoTech Corporation