Whole Genome Design and Modeling for Biomedical & Biotech Applications

Chuan-Hsiung Chang
Institute of Bioinformatics, National Yang-Ming University

cchang@ym.edu.tw

05-02-2006
373 Published Complete Genomes:
- Archaeal: 27 species
- Bacterial: 305 species
- Eukaryal: 41 (Homo sapiens, plants, insects, nematodes, protozoa, fungi, ...)

942 Prokaryotic Ongoing Genomes:
- Archaeal: 55 species
- Bacterial: 887 species
100 times more genomes per year starting two years from now!

Emerging technologies in DNA sequencing

Michael L. Metzker

Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA

Demand for DNA sequence information has never been greater, yet current Sanger technology is too costly, time consuming, and labor intensive to meet this ongoing demand. Applications span numerous research interests, including sequence variation studies, comparative genomics and evolution, forensics, and diagnostics and therapeutics. Several emerging technologies show promise of delivering next-generation solutions to the affordable genome sequencing. In this review we examine some of these next-generation sequencing technologies, single nucleotide addition, and cyclic sequencing, and potential challenges these technologies face in transitioning to high-throughput automation.

Table 1. Companies involved in DNA sequencing technology development

<table>
<thead>
<tr>
<th>Company names</th>
<th>Web site address</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 Life Sciences Corp.</td>
<td><a href="http://www.454.com">www.454.com</a></td>
</tr>
<tr>
<td>Agencourt Biosciences Corp.</td>
<td><a href="http://www.agencourt.com">www.agencourt.com</a></td>
</tr>
<tr>
<td>CE Healthcare, formerly Amersham Biosciences</td>
<td><a href="http://www.amershambiosciences.com">www.amershambiosciences.com</a></td>
</tr>
<tr>
<td>Applied Biosystems, Inc.</td>
<td><a href="http://www.appliedbiosystems.com">www.appliedbiosystems.com</a></td>
</tr>
<tr>
<td>Genovoxx</td>
<td><a href="http://www.genovoxx.de">www.genovoxx.de</a></td>
</tr>
<tr>
<td>Helicos Biosciences Corp.</td>
<td><a href="http://www.helicosbio.com">www.helicosbio.com</a></td>
</tr>
<tr>
<td>LaserGen, Inc.</td>
<td><a href="http://www.lasergen.com">www.lasergen.com</a></td>
</tr>
<tr>
<td>Li-Cor, Inc.</td>
<td><a href="http://www.licor.com">www.licor.com</a></td>
</tr>
<tr>
<td>Microchip Biotechnologies, Inc.</td>
<td><a href="http://www.microchipbiotech.com">www.microchipbiotech.com</a></td>
</tr>
<tr>
<td>Nanofluidics</td>
<td><a href="http://www.nanofluidics.com">www.nanofluidics.com</a></td>
</tr>
<tr>
<td>SeqWright</td>
<td><a href="http://www.seqwright.com">www.seqwright.com</a></td>
</tr>
<tr>
<td>Solexa-Lynx</td>
<td><a href="http://www.solexa.com">www.solexa.com</a></td>
</tr>
<tr>
<td>Visigen Biotechnologies, Inc.</td>
<td><a href="http://www.visigenbio.com">www.visigenbio.com</a></td>
</tr>
</tbody>
</table>
Input data

Genomes & related info

Magic Box

Custom-designed Genomes
Advanced Bioinformatics Core

Core Technology

- Web wrapper agent
  - Retrieve information from Internet
  - Agent learns from user's browsing session
  - Extracted data

- Agent(s)
  - Enable applications to access the Web as if accessing a structured database

Information integration
- Database integration

Analysis (I)
- Web service and workflow

Analysis (II)
- Genomic statistics

Analysis (III)
- Comparative bioinformatics

Service
- Integration of Portal & Grid
Input data

Mining methods

Bioparts & Rules

Computer-Aided Genome Design

Genomes & related info

iCAP

Knowledge Base

GenomeDesigner

Integrated Comparative Analysis Platform for Genomic Data
Our Goal & Approach

Genome Engineering: Genome Design through Genome Comparison

Our Research Interests: How to debug a Bug - Reverse engineering of bacterial genome complexity through genome comparison - to decode the Book of Life
Steps of Genome Analysis

- Organism
  - mRNA
    - Make cDNA
      - Look for EST sequences
  - DNA
    - Genome sequencing & assembly
      - Repeat sequence masking
    - Gene prediction
      - Gene annotation
        - Reconstruction of metabolic pathways & gene regulatory network
          - Comparative genomics
            - Functional genomics
              - Model building & simulation

Genome Design & Engineering

YM-Bioinfo
The Strategy of Bioinformatics

The development and application of global (genome-wide or system-wide) computational approaches to assess gene structures and functions by making use of the information provided by the public genome projects.

The fundamental strategy in a bioinformatics approach is to expand the scope of biological investigation from studying single genes or proteins to studying all genes or proteins, at once, in a systematic and automated fashion.

*Science 278: 601-602, 1997*
The genome is the blueprint that defines an organism and directs every facet of its operation.
Exploring Genomes ➔ The Blueprints for Life

The genome is the blueprint that defines an organism and directs every facet of its operation.
Genotype and Phenotype
To find the rules behind the sequence

Exploring Genomes ➔ The Blueprints for Life

The genome is the blueprint that defines an organism and directs every facet of its operation.

Can we explicitly depict the genome characteristics?

From the genome-wide sequence aspect to the functional implication.
Artificial Life in A Bug Shell
- via Reverse Engineering of Genome Complexity

How to design & build a bacterial genome (the blueprint of life) for a custom-made REAL cell?

- cell size
- generation time
- swimming
- ...

Competition championship for

e.g., gene location, order, strand, operon structure, chromosome structure & number, regulatory circuitry, functional reconstruction & modeling
Currently available approach

The OptStrain Procedure

**Step 1:** Compilation and curation of the Universal database.

**Step 2:** Determination of the maximum yield of the desired product from an optimal substrate choice.

**Step 3:** Minimizing reliance on non-native reactions while satisfying optimal performance criteria.

**Step 4:** Optimal gene deletion determination for coupling biomass production to biochemical formation.

Pharkya et al. (2004), Genome Research, 14(11), 2367-2376.
Our approach

Tools for Bacterial Genome Comparison

- **What** to be compared with?
- **How** to compare them?
  - within one species (different strains)
  - closely-related species
  - moderately-related species
  - distantly-related species
Chromosome comparison of *Vibrio vulnificus* CMCP6 vs. YJ016

*Vibrio vulnificus* CMCP6 chromosome I (3,281,945 bp)

*Vibrio vulnificus* YJ016 chromosome I (3,354,505 bp)

*Vibrio vulnificus* CMCP6 chromosome II (1,844,853 bp)

*Vibrio vulnificus* YJ016 chromosome II (1,857,073 bp)
Chromosome comparison of *Vibrio* species

**Vv** - *Vibrio vulnificus*

**Vp** - *Vibrio parahaemolyticus*

**Vc** - *Vibrio cholerae*
CAGO: a computational system for Comparative Analysis of Genome Organization
Presentation mode for continuous genome features: CURVE

SVG manipulation menu

Reserved circular central blank

Canvas size

Genome features

Legends

Color gradient for presenting feature value

Feature value (min to max)

Feature name

Genome accession number

Mono-color for presenting annotation info.

Features presented in curved mode
Presentation mode for continuous genome features: COLOR GRADIENT

Canvas size

Reserved blank between features

Initial Bend Width

Rule

Grid

Features presented in color gradient mode

Reserved circular central blank
Linear Mode

Canvas size

Disabled SVG manipulation options

Grid

Initial Bend Width

Reserved blank between features

Rule

Linear presentation
Bacterial genomes come in different sizes

NC_000913 4639221 bp Escherichia coli K12, complete genome

NC_000911 3573470 bp Synechocystis sp. PCC 6803, complete genome

NC_000907 1830138 bp Haemophilus influenzae Rd KW20, complete genome

NC_000117 1042519 bp Chlamydia trachomatis D/UW-3/CX, complete genome

NC_000908 580074 bp Mycoplasma genitalium G-37, complete genome

NC_000948 30750 bp Borrelia burgdorferi B31 plasmid cp32-1, complete sequence
CAMP – a computational system for Comparative Analysis of Metabolic Pathways
Metabolic Pathways
Pathway Comparison
Metabolic Profiling
Pathway clustering

Pathway sorting
Species-specific enzymes present in each pathway

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Shared/All Enzymes</th>
<th>V. cholerae</th>
<th>V. vulnificus CMCP6</th>
<th>V. vulnificus YJ016</th>
<th>V. parahaemolyticus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycolysis / Gluconeogenesis</td>
<td>21/43</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle)</td>
<td>11/23</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>• Vc: 4.1.3.6 (VC0797 VC0798 VC0799), 4.1.3.34 (VC0798), 2.8.3.10 (VC0799)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pentose phosphate pathway</td>
<td>18/34</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pentose end glucuronate interconversions</td>
<td>4/54</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>• Vc: 2.7.1.16 (VFA1674) and 5.3.1.4 (VPA1676)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fructose and mannose metabolism</td>
<td>12/65</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>• Vc: 2.7.1.4 (VCA0656)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Vw YJ016: 1.1.1.271 (VV0350), 5.3.1.25 (VV2179), 2.7.1.51 (VV2178)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galactose metabolism</td>
<td>7/38</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>• Vc: 3.2.1.26 (VCA0655)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Vw YJ016: 2.7.1.58 (VVA1574) and 4.2.1.6 (VVA1580)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pyruvate metabolism</td>
<td>26/66</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>• Vc: 4.1.1.- (VC2240)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyoxylate and dicarboxylate metabolism</td>
<td>13/51</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Propanate metabolism</td>
<td>9/40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Butanoate metabolism</td>
<td>16/52</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>• Vc: 4.1.1.5 (VC1589)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carbohydrate Metabolism (10)</td>
<td>96/393</td>
<td>7</td>
<td>0</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>
Enzymes shared in VC and VV YJ016
Gene clustering for functional inference in bacterial genomes

Glycolysis Pathway

Glycolysis Clusters
CICP (Comparative Identification of Conservation Profiles)

Overview

CICP is a computational system for identifying conservation profiles of gene clusters which both have similar chromosomal arrangements and are functionally-coupled in metabolic pathways shared among multiple organisms. Metabolic pathway diagrams and annotations of 157 prokaryotic genomes are obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes). Enzymatic genes involved in a pathway shared by selected organisms and located at neighboring positions on chromosome are grouped together as a gene cluster. With the possibility of genome rearrangement and gene fusion, the presence of each gene and the gene order in each correlated gene cluster are allowed to be different for different organisms.

A conservation profile of the gene clusters shared among multiple organisms is automatically recognized and graphically presented on both chromosome and pathway maps. The profile can be used to investigate evolutionary genome dynamics and improve genome annotation by identifying missing genes.

What can CICP do?

- Identify missing enzymatic genes
- Investigate evolutionary genome dynamics
- Recognize gene fusion events and gene neighborhoods
- Recognize operon rearrangement, transfer, shuffling, and disruption
- Improve genome annotation

Publication of CICP

- CICP poster [abstract] present in ISMB/ECCB 2004

Contact author
If you have any suggestions, comments, or programming problems, please send mail to Hon-Wei Chen: g20123015@ym.edu.tw
CICP computational system

CICP (Comparative Identification of Conservation Profiles)

Select organisms

Bacteria
- Proteobacteria
  - Gamma proteobacteria
  - Beta proteobacteria
  - Delta proteobacteria
  - Alpha proteobacteria
- Firmicutes
- Bacillales
- Lactobacillales
- Clostrida
- Mollicutes
- Actinobacteria
  - Nocardiales
  - Mycobacterium strains
  - Corynebacterium strains
  - Streptomyces strains
  - Thermomyces whippelii strains
- Fusobacteria
- Planococccyes
- Chlamydia
- Spirochete
- Bacteroid
- Cyanobacteria
- Green sulfur bacteria
- Radioresistant bacteria
- Hypervirulent bacteria
- Archaea

User's preferences

Select one metabolic pathway: 1.1.1 Glycolysis / Gluconeogenesis
Select one target organism: No target organism (Identify clusters or profiles only)
Identify all clusters only or make conservation profiles: Identify all clusters only

Display parameters
Align gene clusters: To the genomic position
Zoom: 25% 50% 75% 100% 150% 200%

Clustering parameters
Allowed gene numbers located between two functionally related genes: ≤ 2
Allowed number of reaction steps between two enzymes of your selected metabolic pathway: ≤ 5
Allowed distance (in base pairs) between two functionally related genes: 500 (in base pairs)

Run CICP Reset
Detecting the conservation profiles among all \textbf{Bacillales} strains in terms of the glycolysis pathway.
Search for *Bacillus cereus* based on conservation profiles made in other *Bacillales*. Potential missing enzymatic genes.
Introduction

After annotating a finished genome, we can generally group these predicted genes into three types: first type is known genes which have been functionally characterized; second type is conserved hypothetical genes which are predicted genes that are conserved in many other organisms; third type is hypothetical genes which are predicted genes and not existed in other organisms.

Generally, a newly sequenced genome contains about thirty percents of genes that are annotated poorly or even wrongly. Hence, these poorly characterized hypothetical genes might play an important role in our understanding of life and biology. Given that so little are known about these unknown predicted genes, it is a great challenge for scientists to select a research target worth spending years of time and great amount grants.

Methods outline

After formulating a method to identify potential vital targets from hypothetical proteins, for further study, in vitro experiments have to be taken to validate our candidate proteins.

HyProDB Contents

- **Eukaryotes**
  - Conserved Hypothetical Proteins
  - Hypothetical Proteins

- **Prokaryotes**
  - Conserved Hypothetical Proteins
  - Hypothetical Proteins

Prioritization of hypothetical proteins for functional study
## CARO (Comparative Analysis of Replication Origin)

<table>
<thead>
<tr>
<th>Organism</th>
<th>Accession Number</th>
<th>Replication Origins</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. genitalium</em> (580,074)</td>
<td></td>
<td><img src="image" alt="Graph of M. genitalium" /></td>
</tr>
<tr>
<td><em>M. pneumoniae</em> (816,394)</td>
<td></td>
<td><img src="image" alt="Graph of M. pneumoniae" /></td>
</tr>
<tr>
<td><em>M. gallisepticum</em> (996,422)</td>
<td></td>
<td><img src="image" alt="Graph of M. gallisepticum" /></td>
</tr>
<tr>
<td><em>M. penetrans</em> (1,358,633)</td>
<td></td>
<td><img src="image" alt="Graph of M. penetrans" /></td>
</tr>
<tr>
<td><em>M. pulmonis</em> (963,879)</td>
<td></td>
<td><img src="image" alt="Graph of M. pulmonis" /></td>
</tr>
<tr>
<td>Mycoplasma hominis (232) (892,758)</td>
<td></td>
<td><img src="image" alt="Graph of Mycoplasma hominis" /></td>
</tr>
<tr>
<td><em>U. urealyticum</em> (751,719)</td>
<td></td>
<td><img src="image" alt="Graph of U. urealyticum" /></td>
</tr>
<tr>
<td><em>M. mobile</em> (777,079)</td>
<td></td>
<td><img src="image" alt="Graph of M. mobile" /></td>
</tr>
<tr>
<td>Mycoplasma mycoides (1,211,703)</td>
<td></td>
<td><img src="image" alt="Graph of Mycoplasma mycoides" /></td>
</tr>
<tr>
<td>MF contig 4597 (601,656)</td>
<td></td>
<td><img src="image" alt="Graph of MF contig 4597" /></td>
</tr>
</tbody>
</table>
CATU for Transcription Unit Comparison

-35  -10  RBS  5′UTR  3′UTR  ORF  TSS  Terminator
CAST for Signal Transduction Pathway Comparison

Network elements provide useful design knowledge
Integrated Comparative Analysis Platform (iCAP) for Genomic Data

The component systems:

- **CAGO** (Comparative Analysis of Genome Organization) is a visualization system for comparing various genomic features through intuitive, graphical presentation, including data such as annotation features, nucleotide composition, structural traits, etc.

- **SAGA** (Sequence Atlas Generating Application) can produce varied default genome features and user customized genome characteristics.

- **CAMP** (Comparative Analysis of Metabolic Pathway) uses a systematic method for comparing all the metabolic pathways based on KEGG (Kyoto Encyclopedia of Genes and Genomes) reference pathway data.

- **CICP** (Comparative Identification of Conservation Profiles) is a computational system for identifying conservation profiles of gene clusters which both have similar chromosomal arrangements and are functionally coupled in metabolic pathways shared among multiple organisms.

- **CAST** (Comparative Analysis of Signal Transduction) provides a signal transduction protein database and a tool for comparison of bacterial signal transduction pathways.

- **CATU** (Comparative Analysis of Transcription Unit) is designed to both collect and compare all the transcriptional features of bacterial genes and operons among sequenced genomes.

- **CARO** (Comparative Analysis of Replication Origin) is designed to both collect and compare all the replication origin features of sequenced bacterial genomes.
Welcome to CBS

Comparative Bioinformatic Services

Comparative bioinformatics analysis is the cornerstone of in silico-based approaches to understanding biological systems and processes across species. To provide the infrastructure necessary to sustain biomedical research currently conducted in the NRPGM, especially in projects addressing the needs of comparative bioinformatics analysis, this comparative bioinformatics core service has established this integrated annotation and comparative bioinformatics framework. In this bioinformatics core service project, our Comparative Bioinformatics team of the Advanced Bioinformatics Core (ABC) will highlight the establishment of an integrated comparative bioinformatics infrastructure for genome annotation/re-annotation and comparative analysis. The mission of this project is two-fold. First is to support the current and future comparative bioinformatics needs for biomedical and biotechnological applications, and second is to thoroughly explore the rules behind the genomic sequences. This will not only advance the genomics research in Taiwan, but will also increase Taiwan’s competitiveness in genome-based biomedical and biotechnology industry. The overall goal of this project is to serve the biomedical research community and to promote the delineation of all the genome characteristics through in-depth curatorial genome annotation and cross-species genome comparisons.

CBS Policy

1. This CBS site contains bioinformatics tools and databases which incorporate material contributed or licensed by individuals, companies, or organizations that may be protected by Taiwan and foreign copyright laws. All persons reproducing, redistributing, or making commercial use of this information are expected to adhere to the terms and conditions asserted by the copyright holder.

2. For documents and software available from CBS, we do not warrant or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed. CBS does not endorse or recommend any commercial products, processes, or services.

3. Some CBS web pages may provide links to other Internet sites for the convenience of users. CBS is not responsible for the availability or content of these external sites. nor does CBS endorse, warrant, or guarantee the products, services, or information described or offered at these other Internet sites.

Recommended screen resolution for viewing our CBS web site: 1024 x 768 pixels
Computer-Aided Genome Design

- From templates to knowledge to design