Bioinformatics for (Microbial) Genome Analysis

Yuzhen Ye (yye@indiana.edu)
I609 Bioinformatics Seminar I (Spring 2010)
School of Informatics and Computing
Indiana University
Topics

- **Review**
  - DNA sequencing techniques
  - Genome assembly
  - Gene prediction
  - Function annotation

- **What’s special about analyzing prokaryotic genomes**
  - Assembly may be easier (compact genomes; not so many repeats)
  - Gene finding may be easier (continuous genes)
  - Annotation may be easier (operon structure prediction)
  - HGT (horizontal gene transfer)
## Next generation sequencing (NGS)

<table>
<thead>
<tr>
<th></th>
<th>454 Sequencing</th>
<th>Illumina/Solexa</th>
<th>ABI SOLiD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing Chemistry</td>
<td>Pyrosequencing</td>
<td>Polymerase-based sequence-by-synthesis</td>
<td>Ligation-based sequencing</td>
</tr>
<tr>
<td>Amplification approach</td>
<td>Emulsion PCR</td>
<td>Bridge amplification</td>
<td>Emulsion PCR</td>
</tr>
<tr>
<td>Paired end (PED) separation</td>
<td>3 kb</td>
<td>200-500 bp</td>
<td>3 kb</td>
</tr>
<tr>
<td>Mb per run</td>
<td>100 Mb</td>
<td>1300 Mb</td>
<td>3000 Mb</td>
</tr>
<tr>
<td>Time per PED run</td>
<td>&lt;0.5 day</td>
<td>4 days</td>
<td>5 days</td>
</tr>
<tr>
<td>Read length (update)</td>
<td>250-400 bp</td>
<td>35, 75 and 100 bp</td>
<td>35 and 50 bp</td>
</tr>
<tr>
<td>Cost per run</td>
<td>$ 8,438 USD</td>
<td>$ 8,950 USD</td>
<td>$ 17,447 USD</td>
</tr>
<tr>
<td>Cost per Mb</td>
<td>$ 84.39 USD</td>
<td><strong>$ 5.97 USD</strong></td>
<td><strong>$ 5.81 USD</strong></td>
</tr>
</tbody>
</table>
Illumina HiSeq 2000

- Most recent development; generates up to **200 gigabases** of sequence data and up to two billion paired-end reads on a billion templates per run; $690,000
- gigabase: one billion bases (human genome: 3.2 billion base; *E.coli* K-12 genome size: 4.6 million base)

Next-generation sequencing transforms today's biology

Sanger sequencers

454 sequencer

Next-generation gap

http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.f.268.html
Reads are cheap

>read1
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgac
>read2
gctaagctgggatccgatgacaatgcatgcggctatgctaatggaatggtctttggatccttgagaatagtctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaatgcatgcggctatgcaagctgggatccgatgacaatgcatgcggctatgcaagctgggatccgatgactatgctaatgcatgcggctatgcaagctgggatccgatgac
>read3
tgcggctatgctaatgcatgcggctatgcaagctgggatcttgacgtggtggttgatccttgagaatagtctaatgcatgcggctatgcaagctgggatccgatgacaatgcatgcggctatgcaagctgggatccgatgactatgctaatgcatgcggctatgcaagctgggatccgatgac

......
Genome assembly
Assembly

- **Comparative assembly**
  - comparative (re-sequencing) approaches that use the sequence of a closely related organism as a guide during the assembly process.
- **De novo assembly**
  - reconstructing genomes that are not similar to any organisms previously sequenced
  - proven to be difficult, falling within a class of problems (NP-hard)
  - main strategies: greedy, overlap-layout-consensus, and Eulerian
Fragment assembly: Overlap-Layout-Consensus

**Assemblers:** ARACHNE, PHRAP, CAP, TIGR, CELERA

**Overlap:** find potentially overlapping reads

**Layout:** merge reads into contigs

**Consensus:** derive the DNA sequence and correct read errors
Gaps and contigs

Contig 1

Gap

Contig 2

Filling gap -- up the gaps by further experiments
Mates for ordering the contigs
Read coverage

Assuming uniform distribution of reads:
Length of genomic segment: \( L \)
Number of reads: \( n \)  Coverage \( C = \frac{n l}{L} \)
Length of each read: \( l \)

How much coverage is enough (or what is sufficient oversampling)?
Lander-Waterman model:
\[
P(x) = \left( \lambda^x \cdot e^{-\lambda} \right) / x!
\]
\[
P(x=0) = e^{-\lambda}
\]
where \( \lambda \) is coverage
Repeats complicate fragment assembly

True overlap

Repeat overlap
Gene prediction

aatgcata
gcatgcggctatgctaatgcatgcggctatgctaagctgggata
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgca
agctgggatccgatgactatgctaagctgggatccgatgacaatg
atgcggctatgctaatggaatgggtcttgggatttaccttggaatg
cgaagctgggatccgatgacaatgcatgcggctatgctaatgcatg
cggctatgcaagctgggatccgatgactatgctaagctgggacta
tgctaatgcatusagtggtcttgggatttaccttggaatgctaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgat
gctaatgcatusagtggtcttgggatttaccttggaatgctaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgat
gctaatgcatusagtggtcttgggatttaccttggaatgctaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgat
gctaatgcatusagtggtcttgggatttaccttggaatgctaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgat
gctaatgcatusagtggtcttgggatttaccttggaatgctaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgat
gctaatgcatusagtggtcttgggatttaccttggaatgctaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgat
gctaatgcatusagtggtcttgggatttaccttggaatgctaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgat
gctaatgcatusagtggtcttgggatttaccttggaatgctaatg
What’s a gene

What is a gene, post-ENCODE?
- Gerstein et al. Genome Res. 2007 17: 669-681
- ENCODE consortium: characterization of 1% of the human genome by experimental and computational techniques

Definitions:
- Definition 1970s–1980s: Gene as open reading frame (ORF) sequence pattern
- Definition 1990s–2000s: Annotated genomic entity, enumerated in the databanks
- The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products
Post-ENCODE definition

The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.
Gene prediction is easier in microbial genomes

- Microbial genome tends to be gene rich (80%-90% of the sequence is coding)
- Often no introns
- Highly conserved patterns in the promoter region, transcription and translation start site
Prokaryote gene structure

Transcribed region

start codon

stop codon

Coding region

Untranslated regions

Promoter

Transcription start side

upstream
downstream

Transcription stop side

5’

3’

-k denotes k\textsuperscript{th} base before transcription, +k denotes k\textsuperscript{th} transcribed base
Open reading frame (ORF)

ORF is a sequence of codons which starts with start codon, ends with an end codon and has no end codons in-between.

*Searching for ORFs – consider all 6 possible reading frames: 3 forward and 3 reverse (next slide)*

**Is the ORF a coding sequence?**

1. Must be long enough (roughly 300 bp or more)
2. Should have average amino-acid composition specific for the given organism.
3. Should have codon use specific for the given organism.
Six frame translation of a DNA sequence

- stop codons – TAA, TAG, TGA
- start codons - ATG
Codon usage

- Codon: 3 consecutive nucleotides
- $4^3 = 64$ possible codons
- Genetic code is degenerative and redundant
  - Includes start and stop codons
  - An amino acid may be coded by more than one codon (degeneracy & wobbling pairing)
  - Certain codons are more in use
- Uneven use of the codons may characterize a real gene
Codon frequency

Input sequence

- frequency in coding region
- frequency in non-coding region

Compare

Coding region or non-coding region
A simple calculation assuming independence of nucleotides

<table>
<thead>
<tr>
<th>codon position</th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28%</td>
<td>33%</td>
<td>18%</td>
<td>21%</td>
</tr>
<tr>
<td>2</td>
<td>32%</td>
<td>16%</td>
<td>21%</td>
<td>32%</td>
</tr>
<tr>
<td>3</td>
<td>33%</td>
<td>15%</td>
<td>14%</td>
<td>38%</td>
</tr>
<tr>
<td>frequency in genome</td>
<td>31%</td>
<td>18%</td>
<td>19%</td>
<td>31%</td>
</tr>
</tbody>
</table>

\[
\frac{P(x|ORF)}{P(x|\text{random})} = \prod_i \frac{P(A_i \text{ at ith position})}{P(A_i \text{ in the sequence})}
\]

Score of AAAGAT:

\[
\]
Coding region prediction using hexmer frequencies

Coding potential – hexmer frequencies in coding versus in non-coding regions $\sum \log \left( \frac{C_i(X)}{N_i(X)} \right)$
Simple HMM for prokaryotic genes

<table>
<thead>
<tr>
<th>States</th>
<th>Transition probability</th>
<th>Emission probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position 1</td>
<td>1</td>
<td>1-q</td>
</tr>
<tr>
<td>Position 2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Position 3</td>
<td></td>
<td>1-p</td>
</tr>
</tbody>
</table>

Non-Coding Region

A: $p_A$
C: $p_C$
G: $p_G$
T: $p_T$
HMM based gene predictors for microbial genomes

- **GenMark** [Borodovsky, McInnich – 1993, Comp. Chem., 17, 123-133] 5th order HMM
Additional information for gene prediction

- Upstream regions of genes often contain motifs that can be used for gene prediction
Promoter structure in prokaryotes

Transcription starts at offset 0.

- Pribnow Box (-10)
- Gilbert Box (-30)
- Ribosomal Binding Site (+10)
Ribosomal binding Site

1055 E. coli Ribosome binding sites listed in the Miller book
Why we need to do function annotation?

Fig from: Network-based prediction of protein function. Molecular Systems Biology 3:88. 2007
What’s function?

- The definition of biological function is ambiguous (context dependent)
  - FOXP2 is involved in human-specific transcriptional regulation of CNS development
  - the transcription factor FOXP2 (forkhead box P2) is the only gene implicated in Mendelian forms of human speech and language dysfunction
  - two human-specific amino acids alter FOXP2 function by conferring differential transcriptional regulation in vitro…
  - Nature 462, 213-217, 2009

- It is obvious that the biological function of a protein has more than one aspect
How to describe function?

- in a computationally amenable way?
- Human language
- **Controlled** vocabulary
  - EC (Enzyme Commission Classification)
    1. -.- Oxidoreductases.
    1. 1. -.- Acting on the CH-OH group of donors.
    1. 1. 1.- With NAD(+) or NADP(+) as acceptor.
    1.1.1.1 Alcohol dehydrogenase.
    1.1.1.3 Homoserine dehydrogenase.
  - GO (Gene Ontology)
    - http://www.geneontology.org
What information can be used for function annotation?

- **Sequence based approaches**
  - Protein A has function X, and protein B is a homolog (ortholog) of protein A; Hence B has function X

- **Structure-based approaches**
  - Protein A has structure X, and X has so-so structural features; Hence A’s function sites are …. 

- **Motif-based approaches (sequence motifs, 3D motifs)**
  - A group of genes have function X and they all have motif Y; protein A has motif Y; Hence protein A’s function might be related to X

- **“Guilt-by-association”**
  - Gene A has function X and gene B is often “associated” with gene A, B might have function related to X
    - Associations
      - Domain fusion, phylogenetic profiling, PPI, etc.

- **Meta-approaches**
Operon

- An operon is a group of genes (adjacent to one another on the chromosome) that are **transcribed at the same time**.
- They usually control an important biochemical process.
- Examples: Tryptophan operon & Lac operon
- Present in Prokaryotes, but not in Eukaryotes
The tryptophan repressor in Bacteria

E. coli chromosome

mRNA molecule

enzymes for tryptophan biosynthesis

promoter

operator

inactive repressor

RNA polymerase

tryptophan

active repressor

GENES ARE ON

GENES ARE OFF

(negative control)

Figure 7-34 Molecular Biology of the Cell (© Garland Science 2008)
Tryptophan repressor is a member of Helix-turn-helix family

The binding of tryptophan cause conformation change

GENES ARE ON

GENES ARE OFF

Figure 7-36  *Molecular Biology of the Cell* (© Garland Science 2008)
Some bacterial gene regulatory proteins can act as either a repressor or an activator, depending on the precise placement of their DNA-binding sites.
Dual control of the $\text{Lac}$ operon

Activator: CAP (catabolite activator protein)

Repressor: $\text{Lac}$ repressor

Figure 7-39 Molecular Biology of the Cell (© Garland Science 2008)
Operon predictor

- Two basic types of approaches
  - identifying operons that are conserved in multiple species, as genes that remain adjacent across long stretches of evolutionary time are likely to be in the same operon (problem: evolutionarily new operons)
  - probabilistic ‘distance models’: genes in the same operon tend to be separated by fewer base pairs of DNA
    \[ p(d|\text{operon}) \Rightarrow p(\text{operon}|d) \]

- MicrobesOnline Operon Predictions (http://www.microbesonline.org/operons/) are based on a method that uses both comparative and distance information, and infers a genome-specific distance model from preliminary comparative-only predictions
Horizontal Gene Transfer

- Also called Lateral Gene Transfer; HGT and LGT for short

- 3 ways to do it
  - **Transformation**- naked DNA, short pieces, common in bacteria that transform
  - Transduction – phage, donor/recipient share receptors, closely related bacteria, DNA: amount in phage head
  - Conjugation-plasmids/transposons, cell to cell contact, distant relations, long DNA
Requirements for Transfer

- Proximity to donor DNA
- Stability of DNA in environment
- Vector transmission
- Uptake and insertion
- Maintenance
- Stabilization
- Selection

What Limits/Prevents Transfer

- Instability in new host
- Restriction systems
- GC/Codon usage incompatibility
- Splicing and other signals incorrect
- RNA editing
- Lack of appropriate interacting genes
The question: how much HGT really goes on?

- Rare going from prokaryotes into multicelled euks because must go into egg/sperm
- Exception is mitochondria, plastids
- Instances of Euks -> Proks ?!?!
- How often does it happen in proks?
LGT and the nature of bacterial innovation
Ochman, Lawrence, and Groisman, Nature 405:299-304

- Single celled organism, genome varies only by an order of magnitude.
- Narrow taxonomic groups, phenotypic diversity is remarkable.
- Usually have a unique set of physiological characters to define its particular ecological niche.
<table>
<thead>
<tr>
<th>Bacterial Species</th>
<th>Utilize Lactose</th>
<th>Utilize Citrate</th>
<th>Produce H₂S</th>
<th>Produce Indole</th>
<th>Produce Urease</th>
<th>Lysine Decarboxylase</th>
<th>Lifestyle</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Escherichia coli</em></td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>Mammalian commensal</td>
</tr>
<tr>
<td><em>Shigella flexneri</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>Primate pathogen</td>
</tr>
<tr>
<td><em>Salmonella enterica</em></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>Mammalian pathogen</td>
</tr>
<tr>
<td><em>Klebsiella pneumoniae</em></td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>Soil</td>
</tr>
<tr>
<td><em>Serratia marcescens</em></td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>Soil</td>
</tr>
<tr>
<td><em>Yersinia pestis</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Mammalian pathogen</td>
</tr>
</tbody>
</table>
How can you detect HGTs?

- DNA sequence information
  - Phylogenetic trees
  - G+C Content
  - Codon bias

- Sequences new to a genome will retain (for a while) the signatures of the donor genome and distinguished from ancestral DNA
Comparing a gene tree with the rRNA tree
G+C Content & Codon Usage

- DNA is double stranded, G pairing with C
- Measure the amount of G+C content in regions
- If one region varies from most of the genome, than likely HGT

52%  47%  52%
Genomic island predictors

- Utilize structures of genomic islands
- Utilize sequence composition characteristics (such as G+C ratio and dinucleotide)
Bacterial pathogenomics
Mark J. Pallen & Brendan W. Wren
Nature 449, 835-842(18 October 2007)