Bioinformatics
Maxam-Gilbert sequencing


G reaction   G+A reaction   T+C reaction   C reaction

G   G+A   T+C   C

3'

*ACTAGCTAG
*ACTAGCTA
*ACTAGCT
*ACTAGC
*ACTAG
*ACTA
*ACT
*AC
*A

5'

Figure 5.33
Determining DNA Base Sequence

• Restriction enzymes cut DNA at specific sites
  – Normally protect bacteria from viral DNA
  – Enzyme cut reveals location of specific sequence

• PCR uses short oligonucleotides
  – Primers bind complementary sites
  – Archaeal enzyme synthesizes DNA
  – Rapidly amplifies segment of DNA

• Sanger method determines sequence of up to 1000 bases
Dideoxynucleotide DNA sequencing—the chain termination method

Figure 5.35
Whole genome shotgun sequencing


Figure 5.39
Whole-genome Sequencing

- Break genome into thousands of pieces
- Determine sequence of many short pieces
- Computer determines sequence overlap to recreate entire genome sequence.
Fiber optic slides for DNA sequencing

(a) Loading capture beads onto the fiber-optic slide

(b) Schematic showing the location of the fiber-optic slide in the GS FLX instrument

Figure 5.42

Courtesy of 454 Sequencing. © 2010 Roche Diagnostics.

Table 12.2  Gene function in bacterial genomes

<table>
<thead>
<tr>
<th>Functional categories</th>
<th>Escherichia coli (4.64 Mbp)$^a$</th>
<th>Haemophilus influenzae (1.83 Mbp)$^a$</th>
<th>Mycoplasma genitalium (0.58 Mbp)$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism</td>
<td>21.0</td>
<td>19.0</td>
<td>14.6</td>
</tr>
<tr>
<td>Structure</td>
<td>5.5</td>
<td>4.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Transport</td>
<td>10.0</td>
<td>7.0</td>
<td>7.3</td>
</tr>
<tr>
<td>Regulation</td>
<td>8.5</td>
<td>6.6</td>
<td>6.0</td>
</tr>
<tr>
<td>Translation</td>
<td>4.5</td>
<td>8.0</td>
<td>21.6</td>
</tr>
<tr>
<td>Transcription</td>
<td>1.3</td>
<td>1.5</td>
<td>2.6</td>
</tr>
<tr>
<td>Replication</td>
<td>2.7</td>
<td>4.9</td>
<td>6.8</td>
</tr>
<tr>
<td>Other, known</td>
<td>8.5</td>
<td>5.2</td>
<td>5.8</td>
</tr>
<tr>
<td>Unknown</td>
<td>38.1</td>
<td>43.0</td>
<td>32.0</td>
</tr>
</tbody>
</table>

$^a$Chromosome size, in megabase pairs. Each organism listed contains only a single circular chromosome.
Figure 12.6

The figure shows the percent of genes in certain functional categories for Bacteria and Archaea. The categories include Carbohydrate metabolism, Cell membrane, Coenzyme metabolism, Energy production, Unknown function, and General prediction.

- Carbohydrate metabolism: Bacteria has 4% genes, Archaea has 2%.
- Cell membrane: Bacteria has 6% genes, Archaea has 4%.
- Coenzyme metabolism: Bacteria has 2% genes, Archaea has 2%.
- Energy production: Bacteria has 8% genes, Archaea has 12%.
- Unknown function: Bacteria has 10% genes, Archaea has 14%.
- General prediction: Bacteria has 12% genes, Archaea has 14%.

The bars for Bacteria are generally lower than those for Archaea, with the exception of Energy production and Unknown function, where Bacteria have slightly higher percentages.

The y-axis represents the percent of genes, ranging from 0 to 14%.
Making a DNA chip

Figure 5.46a
(a)

M1
Shine light
M1
Couple
M2
Repeat

Figure 5.46b
(b)

Touch surface
Move pins
Repeat

Synthesize short ss oligonucleotides complementary to genes X, Y, and Z.

Affix DNA to chip at known locations.

Probe chip with labeled mRNA and scan chip.

Growth condition 1:
- Gene X expressed
- Genes Y and Z not expressed

Growth condition 2:
- Gene X not expressed
- Genes Y and Z expressed
Using a DNA microarray
Figure 12.11

$M_r$ (kDa)

pH

160
81
43
25
12
7
6
5

Jack Parker
Figure 12.13

Chemotaxis proteins
Flagellar/motility
MCP proteins
Not motility related
Figure 12.14

- Sample
- Flight tube
- Laser beam
- TOF detector
- Matrix
- Ions
- Charged grid
- Mass ($m/z$)
- Intensity
Figure 12.15

Paralogs

Paralogs

Paralogs

Paralogs

Orthologs

Orthologs

Species 1

Species 2

Species 1

Species 2

Gene A

Gene B

Ancestral gene in ancestral species

Divergence of species

Gene duplication
Figure 12.17

Prophage

PAI II

E. coli strain Genome (bp)
K-12 4,639,221
536 4,938,875
073 5,231,428
Figure 12.18

Chromosomal islands

Plasmids

Transposon

Pathogenicity island

Integrated phage DNA
The human genome sequence

- ~2500 protein coding genes
- DNA that encodes protein coding genes accounts for only ~2% of genome
- Some chromosomes are more gene-rich than others
- There are several thousand non-protein coding genes that code for RNAs
- ~50% of protein coding genes have an unknown function
- The human genome is almost identical (99.9%) identical in all people