MICROBIAL GENETICS – GENOMICS

Genomics, Functional Genomics, Structural Genomics

Homology

Sequencing Whole Genomes
   Library of Cloned Fragments
   Contigs
   Annotation
   Orthologous, Paralogous

Global Studies of Gene Expression
   DNA arrays/Gene arrays/Microarrays/DNA chips
   Making arrays
   Labelling RNAs (by making cDNA copies)
   Hybridization of cDNAs to microarrays

Concept of Core Gene Pool and Flexible Gene Pool

   Core Gene Pool
      Replication, Transcription, Translation, Glycolysis, Cell Wall Synthesis

   Flexible Gene Pool
      Insertion sequences, transposons, integrons, plasmids, phages (prophages), genomic islands

      Pathogenicity, Drug Resistance, Toxin Synthesis/Secretion, Conjugation

Proteomics
GENOMICS - STUDY OF ALL GENES AND SITES WITHIN THE CHROMOSOME(S) OF AN ORGANISM BY DETERMINING AND ANALYSING THE NUCLEOTIDE SEQUENCE OF THE ENTIRE CHROMOSOME(S)

FUNCTIONAL GENOMICS - STUDY OF FUNCTIONS OF A CELL OR ORGANISM BY ANALYSIS OF CHROMOSOME NUCLEOTIDE SEQUENCE

STRUCTURAL GENOMICS - STUDY OF 3D SHAPES OF PROTEINS BY COMPARISON OF AMINO ACID SEQUENCES [DEDUCED FROM GENE SEQUENCES] TO PROTEIN STRUCTURE DATA BASES

BIOINFORMATICS - DEVELOPMENT OF SOFTWARE AND ALGORITHMS TO ANALYZE GENE AND PROTEIN SEQUENCES AND TO COMPARE THEM TO DATA BASES

ALGORITHM - RULES FOR PROCEDURES FOR SOLVING MATHEMATICAL PROBLEM, USUALLY INVOLVING REPETITIVE OPERATIONS

HOMOLOGY - EXTENT OF SIMILARITY [OR IDENTITY] BETWEEN GENES OR PROTEINS. PROTEINS WHICH PERFORM THE SAME FUNCTIONS OFTEN HAVE SIMILAR AMINO ACID SEQUENCES. GENES WHICH ENCODE SIMILAR PROTEINS HAVE SIMILAR NUCLEOTIDE SEQUENCES. SOME PROTEINS [AND GENES] ARE VERY SIMILAR AND ARE SAID TO HAVE A HIGH DEGREE OF HOMOLOGY. OTHER PROTEINS AND GENES ARE MORE DISTANTLY RELATED AND SHARE A LOW DEGREE OF HOMOLOGY. UNRELATED PROTEINS HAVE NO SIMILARITY IN SEQUENCE.

DNA SEQUENCING - CLONE GENE OR FRAGMENT OF INTEREST, PRIMER, SYNTHESIS OF LABELLED DNA AS OVERLAPPING FRAGMENTS, EACH INCREASING IN LENGTH BY ONE NUCLEOTIDE. DETERMINATION OF LAST NUCLEOTIDE ON 3' END OF EACH FRAGMENT GIVES ORDER OF NUCLEOTIDES POLYMERIZED BY THE NEW SYNTHESIS AND THE NUCLEOTIDE SEQUENCE ALONG ONE STRAND. USE ANOTHER PRIMER TO CONFIRM SEQUENCE OF COMPLEMENTARY STRAND.

WHOLE GENOME RANDOM SEQUENCING -
WHOLE GENOME SHOTGUN SEQUENCING
WHOLE GENOME - RANDOM SEQUENCING
- SHOTGUN SEQUENCING

- PREPARE LIBRARY OF RANDOM FRAGMENTS
  - SMALL + LARGE INSERTS
  - 5-10X REDUNDANT (TO HAVE AT LEAST ONE COPY OF EVERY REGION OF CHROMOSOME)

- EXTRACT CHROMOSOME
  - SHEAR TO ~2, ~20, ~50 KB
  - BLUNT ENDS

- CLONE
  - SHORT INSERTS - PLASMIDS
  - LONG INSERTS - PHAGE LAMBDA 20 KB
  - COSMIDS 50 KB

NEED ~20,000 CLONES FOR 2 MEGA BP CHROMOSOME
SEQUENCE

~20,000 SEQUENCES

~500 BP/SEQUENCE

→ PRIMERS←

△ COMPILE
FIND OVERLAPPING REGIONS

SEQ 1

5'-----3'

2
3
4

OVERLAPPING REGIONS
CLOSE GAPS BETWEEN CONTIGS
EX: WALK ALONG LARGE INSERTS
BY SYNTHESIZING PRIMERS
COMPLEMENTARY TO
CHROMOSOME INSERTS +
USING THEM TO PRIME
ADDITIONAL SEQUENCING

ANNOTATION
FIND PROTEIN CODING REGIONS-
OPEN READING FRAMES - ORF
80 CONTIGUOUS CODONS
WITH START + STOP CODONS
+ SHINE DALGARNO
Identify genes (CDS) by comparing amino acid sequence against all known protein sequences.

Identify tRNA + rRNA genes + operons.

Identify repeat sequences e.g., insertion sequences.

Homology = degree of identity [amino acids + nucleotides].

Orthologous proteins - similar proteins from different species - orthologs.

Paralogous proteins - similar proteins within same organism - paralogs.
<table>
<thead>
<tr>
<th>Organism</th>
<th>Chromosome (BP)</th>
<th>Total Genes</th>
<th>Known</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>4.6x10^6</td>
<td>5295</td>
<td></td>
<td>2936</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>1.8x10^6</td>
<td>1738</td>
<td>1473</td>
<td></td>
</tr>
<tr>
<td>Helicobacter pylori</td>
<td>1.7x10^6</td>
<td>1589</td>
<td>1068</td>
<td></td>
</tr>
<tr>
<td>Methanococcus Jannaschii</td>
<td>1.7x10^6</td>
<td>1727</td>
<td>1232</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Haemophilus influenzae</td>
<td>Methanococcus jannaschii</td>
<td>Helicobacter pylori</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------------</td>
<td>--------------------------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td><strong>GENOME (x 10^6 BP)</strong></td>
<td>1.8</td>
<td>1.7</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td><strong>R RNA OPERONS</strong></td>
<td>6</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>T RNA GENES</strong></td>
<td>58</td>
<td>37</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td><strong>ORFs</strong></td>
<td>1,738</td>
<td>1,727</td>
<td>1,589</td>
<td></td>
</tr>
<tr>
<td><strong>UNKNOWN FUNCTION</strong></td>
<td>616</td>
<td>1,089</td>
<td>707</td>
<td></td>
</tr>
<tr>
<td><strong>ASSIGNED FUNCTION</strong></td>
<td>1,122</td>
<td>638</td>
<td>882</td>
<td></td>
</tr>
</tbody>
</table>

**NUMBER OF GENES FCR:**

<table>
<thead>
<tr>
<th>Category</th>
<th>%a</th>
<th>%b</th>
<th>%c</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMINO ACID METABOL</td>
<td>71</td>
<td>64</td>
<td>42</td>
</tr>
<tr>
<td>COENZYME SYNTHESIS</td>
<td>56</td>
<td>49</td>
<td>57</td>
</tr>
<tr>
<td>CELL ENVELOPE</td>
<td>102</td>
<td>24</td>
<td>101</td>
</tr>
<tr>
<td>CELL DIVISION, ETC</td>
<td>67</td>
<td>26</td>
<td>125</td>
</tr>
<tr>
<td>SUGAR/NITROGEN METABOLISM</td>
<td>43</td>
<td>19</td>
<td>24</td>
</tr>
<tr>
<td>DNA METABOLISM</td>
<td>91</td>
<td>53</td>
<td>90</td>
</tr>
<tr>
<td>ENERGY METABOLIS</td>
<td>143</td>
<td>158</td>
<td>100</td>
</tr>
<tr>
<td>LIPID/PHOSPHOLIPID</td>
<td>34</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>OTHER CATEGORIES</td>
<td>19</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>PROTEIN FATE (FOLDING, ETC)</td>
<td>71</td>
<td>38</td>
<td>41</td>
</tr>
<tr>
<td>PROTEIN SYNTHESIS</td>
<td>116</td>
<td>117</td>
<td>99</td>
</tr>
<tr>
<td>NUCLEIC ACID PRECURSORS</td>
<td>50</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>REGULATORS</td>
<td>64</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>TRANSCRIPTION</td>
<td>30</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>TRANSPORT</td>
<td>165</td>
<td>56</td>
<td>88</td>
</tr>
<tr>
<td>HYPOTHETICAL, CONSERVED</td>
<td>351</td>
<td>547</td>
<td>186</td>
</tr>
<tr>
<td>HYPOTHETICAL, NOT CONSERVED</td>
<td>237</td>
<td>513</td>
<td>497</td>
</tr>
<tr>
<td>OTHER UNKNOWN</td>
<td>28</td>
<td>29</td>
<td>24</td>
</tr>
</tbody>
</table>

a BASED ON HOMOLOGY OF DEDUCED PROTEIN SEQUENCE TO PREVIOUSLY CHARACTERIZED PROTEINS
b PERCENTAGES (ROUNDED OFF) OF TOTAL GENES
c AUTOTROPHIC ARCHAEON, GROWS ON THE OCEAN FLOOR AT BASE OF 2800 METER VOLCANO, 200 ATMOSPHERES PRESSURE, TEMP RANGE OF 48 TO 94°C, WITH OPTIMUM OF 85°C, STRICT ANAEROBE, PRODUCES METHANE BY OXIDIZING H₂ AND REDUCING CO₂.
d PLUS TWO PLASMIDS OF 58 AND 16 KBP
e IN OTHER CREATURES
FUNCTIONAL GENOMICS INCLUDES GLOBAL STUDY OF GENE EXPRESSION

DNA ARRAYS/GENE ARRAYS/MICROARRAYS/DNA CHIPS

= DNA FRAGMENTS CORRESPONDING TO EVERY GENE (ORF) FIXED TO SEPARATE, TINY SPOTS ON A SLIDE

USED TO DETECT + QUANTITIZE mRNA FROM EVERY GENE

STEP 1

- MAKING THE ARRAY
  Δ EXTRACT CHROMOSOMAL DNA

Δ AMPLIFY FRAGMENTS BY PCR CORRESPONDING TO EACH ORF (REQUIRES SPECIFIC PRIMER PAIRS)
**STEP 2**

- Extract RNA from culture(s)
  - E.g., Culture - aerobic
  - Culture - anaerobic

- Synthesize cDNA from each mRNA by reverse transcriptase; labelling:
  - Aerobic culture cDNA: dCTP - red fluorescent tag
  - Anaerobic culture cDNA: dCTP - green fluorescent tag

Affix each fragment to separate spot on slide
HYBRIDIZE CDNA's TO MICROARRAY TO QUANTITATE CDNA

- MIX RED- AND GREEN-TAGGED CDNA's

- ANNEAL (HYBRIDIZE) TO MICROARRAY

- REMEMBER - DNAs ON SLIDE ARE IN EXCESS (OVER CDNA's)

- MEASURE RED + GREEN FLOURESCENCE

- FLOURESCENCE INTENSITY IS PROPORTIONAL TO CONCENTRATION OF EACH CDNA

- CDNA AMOUNT REFLECTS AMOUNT OF mRNA OF EACH GENE (ORF)

- CORRELATE EXPRESSION LEVEL OF GENES WITH GROWTH CONDITION
Microarray before annealing:

Slide

- Immobilized DNA fragments corresponding to ORF's 1...4020 before annealing to labelled cDNA

After annealing:

- ORF's expressed in aerobic growth
- ORF's expressed in anaerobic growth
- ORF's expressed in aerobic + anaerobic growth
**MICROARRAYS TO ANALYZE GLOBAL GENE EXPRESSION IN BACILLUS SUBTILIS GROWN IN AEROBIC AND ANAEROBIC CONDITIONS**

- **4.2 \times 10^6** bp genome (4.2 mbp genome)  
  - 4100 ORFs  
  - 4020 ORFs SYNTHESIZED AND IMMOBILIZED

**Aerobic Growth** - Yeast Extract - Tryptone - Glucose with vigorous shaking to mix oxygen into the liquid medium.
- cDNAs labelled with red fluorescent dye

**Anaerobic Growth** - The same medium plus nitrate or nitrite as electron acceptor, the culture bottle flushed with inert gas.
- cDNAs labelled with green fluorescent dye

<table>
<thead>
<tr>
<th>Genes involved in nitrate/nitrite reduction</th>
<th>Level of induction (or repression)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitrate Reductase</td>
<td>100 - 600 fold</td>
</tr>
<tr>
<td>Three subunits</td>
<td></td>
</tr>
<tr>
<td>Assembly factor</td>
<td>200</td>
</tr>
<tr>
<td>Nitrite Reductase</td>
<td>10 - 40</td>
</tr>
<tr>
<td>Two subunits</td>
<td></td>
</tr>
<tr>
<td>Global Anaerobic Regulator</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genes involved in electron transport</th>
<th>2 - 20 fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 genes</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unknown genes induced</th>
<th>2 - 15 fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 genes</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unknown genes repressed</th>
<th>(2 - 7 fold repressed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 genes</td>
<td></td>
</tr>
</tbody>
</table>
Major result of genomics studies of prokaryotes

Core gene pool
"Chromosome (for each species)

Flexible gene pool
Genomic islands
Phages (prophage)
Plasmids
Integrons
Transposons
Insertion sequences

Pathogenicity
Drug resistance
Toxin/secretion
Conjugation

Replication
Transcription
Translation
Glycolysis
Cell wall
- Proteomics - Understanding everything about all proteins of given organism

  How many?
  Function of each?
  Which other proteins do they interact with?

- Structural genomics includes:
  Structural biology, genomics, + computational modeling

  Goal - to identify all proteins of an organism - high throughput protein purification crystallization structure determination correlation domains + regions of proteins with function