The Genome and Biotechnologies

Pyrosequencing  Genome Analyzer  SOLiD
454  Reversible Terminator Chemistry  Ligation-based extension
Outline

• Basic Molecular Biology
  – Genes, Proteins, etc.
  – Genomes
  – Central Dogma

• Observing the Central Dogma
  – PCR - amplification
  – DNA Sequencing - genome
  – Microarrays - transcriptome
  – ChIP – Protein DNA interactions
  – Yeast Two-Hybrid – Protein Protein Interacts.
Preliminaries

• Life is survival of information
• Properties of life:
  – Information exchange (communication)
  – Procreation (passing on information)
  – Evolution (change)

- A machine that’s set in motion and never stops
Preliminaries

- Top-down Organization of life
  - Social groups etc.
  - Organisms
  - Species, etc.
  - Organs, Tissues
  - Cells: units of life
  - Organelles etc.
  - Molecules of life: DNA, RNA, proteins
Preliminaries

- **DNA**: Inheritable information
- Units of Inheritance: genes, regulatory regions, ?
- **Proteins**: Day-to-day footwork
- Both are complex polymer molecules
  - DNA: String over the alphabet \{A,C,G,T\}
  - A,T and C,G are complementary bases
Genes to Proteins

<table>
<thead>
<tr>
<th>First Base of Codon</th>
<th>Second Base of Codon</th>
<th>Third Base of Codon</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>U</td>
<td>UCAG</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>UCAG</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>UCAG</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>UCAG</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>UCAG</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>UCAG</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>UCAG</td>
</tr>
</tbody>
</table>

- **U**: phe, phe, leu, leu, leu, leu, ile, ile, ile, ile, met*, val, val, val
- **C**: UCU, UCC, UCA, UCG, CCU, CCC, CCA, CGC, ACA, ACC, ACC, ACA
- **A**: UAU, UAC, UAA, UAG, CAU, CAC, CAA, CAG
- **G**: UGU, UGC, UGA, UGG

- **Stop Codons**: UAA, UAG, UGA
- **Start Codon**: AUG

- **Amino Acids**:
  - phe: phenylalanine
  - ser: serine
  - tyr: tyrosine
  - cys: cysteine
  - pro: proline
  - his: histidine
  - arg: arginine
  - asp: aspartic acid
  - asn: asparagine
  - ser: serine
  - gln: glutamine
  - gln: glutamine
  - met*: methionine
  - val: valine
Genes to Proteins

Central Dogma: DNA $\Rightarrow$ RNA $\Rightarrow$ Proteins

Biological Information Flow
Genomes

Organization and complexity

- Genomes are the union of all DNA in an organism (there are different types of DNA: nuclear and mitochondrial)
- Only small % (2%) of the human genome is genes. The rest contains various promoter regions and “junk?” (>50%)
- Genome sizes vary among organisms, shortest for Phages and Viruses, longest for mammals and some plants (figure from Baldi)
Evolution

- Changes in the genomes

- Mutations: changes in genome driven by random or particular events. Can be single base change or larger events.

- Recombination: mixing of genomes to produce a new one

- Natural selection: beneficial changes are passed on
Similarity of genomes (i.e. organisms)

- Evolution implies that different organisms would have common ancestors
- Thus similarity comparisons (homology searches) provide clues to evolutionary ancestry (phylogeny)
Central Dogma of Molecular Biology
Transcriptional Regulation of Gene Expression
Transcriptome

All possible gene expressions in the organism
Organization and Complexity

- Transcriptome is the measurable level of all different mRNA’s in an organism
- One DNA template multiple mRNAs: alternative splicing
- DNA to mRNA: one way street because of alternative splicing
- The “when and where” of mRNA concentration is coded in the promoter regions, and possibly elsewhere

Evolution

- Evolution of gene expression under emergent properties like network organization

Similarity of Organisms

- Comparison of gene expression from a “system’s perspective”
Proteome

Localization, abundance, and interaction of all proteins in an organism

- Structure: Amino acid sequence, 3D crystal structure
- Structure => Function?
- Sequence homology not always good indicator of functional similarity
- Study of protein expression
Other -omes

- Interactome
- Cellome
- Biome
- Metabolome
- Envirome
- ?
BioTechnologies

• Observing the Central Dogma
  – seeing the minute: PCR
  – DNA growing and sequencing
  – gene expression: microarrays
  – DNA-protein interactions: immunoprecipitation
  – protein-protein interactions: yeast 2 hybrid

• Large-Scale Technologies:
  – Thousands of measured variables
  – Require computational processing
1. Seeing the minute: PCR

Producing multiple copies of given DNA fragment (amplification)

1. Start with a double stranded DNA molecule
   • Separate strands into templates by heating the mixture
   • Cool to allow “primers” to attach to single strands
   • The primers identify the starting points for DNA synthesis
   • DNA synthesis of strands complementary to the templates
   • Go to 1.

After cycle 30, >1 billion identical molecules ($2^{30} = 1.07 	imes 10^9$)
PCR properties

• The primers can determine the amplified DNA fragment if chosen to flank that region
• n steps of the above produce $2^n$ copies of the intended DNA fragment

$2^{30} \sim 10^9$
Challenges

- Primer Design: short and unique
- Number of cycles
- Signal sensitivity
2. Growing DNA: Synthesizers

ABI 3900 High-Throughput DNA Synthesizer
3. DNA Sequencing

Reading the string: determining the exact positions of the base pairs A, C, G, T

- Break DNA into manageable fragments (500 – 700 bp)
- Sequence the fragments
- Assemble/map the fragments
Sequencing approaches

- Shotgun sequencing: “random” overlapping fragments (Celera)
- Mapped sequencing: shorter sequences are anchored (Human Genome Consortium)
- Techniques: Sanger, pyrosequencing, etc. (454, Illumina, Polonies)
Fragment assembly and coverage
Challenges

- Obtaining clean DNA
- Read quality
- Uniform coverage
- Assembly/mapping of DNA
4. Microarrays

- Testing for the presence and concentration of mRNAs, Proteins, or Protein-DNA interactions
- Can do $10^6$ experiments at a time!
- Hypothesis generation vs. promise of complete description on a large scale
What are Microarrays good for?

• Gene Expression Profiling
  - Genes that behave differently to treatments in same organisms
  - Different organisms
• Identifying naturally oscillating genes in the cell: example cell cycling genes in yeast
• Identifying SNPs
• Tumor vs normal cells
• Transcription Factor Regulation
Microarray Formats

Spotted Microarrays

• Ed Southern 25 years ago, Patrick Brown recently
• Glass slide DNA arrays: 100,000 sites per 1cm$^2$

Oligonucleotide arrays

• Photolithographic method (Affymetrix Inc.) just like computer chips
• Masks used to synthesize oligonucleotides to a chip 1,000,000 sites per 1cm$^2$. 
Two-color arrays vs one

Diagram showing the process of gene expression analysis:
1. RNA Isolation
2. Reverse Transcriptase Labeling
3. "Red Fluorescent" Probes
4. "Green Fluorescent" Probes
5. Combine Targets
6. Hybridize to Microarray

Images of GeneChip devices are also shown.
Other Microarray Technologies

- Photolytography
- Ink jet (Agilent),
- Addressable beads (Lynx),
- etc.
Sources of Error in Microarrays

- Length of probes
- Cross and self hybridization
- Environmental conditions
- Array (plate) and dye idiosyncrasies
- Other?
Challenges

• Probe design
• Plate design
• Data Analysis:
  – over/under expression
  – classification,
  – clustering,
  – regulation inference,
  – gene networks
5. ChIP

ChIP: Chromatin Immuno-Precipitation
Resolves: DNA-Protein Interactions

- Protein bound to DNA
- Sheared DNA
- Antibody binding and protein release
- PCR the isolated DNA
- Sequence the fragments
Challenges

- Identifying anti-bodies

- Identifying precise location of binding

- Distinguishing signal from noise
6. PPI: Two-Hybrid Screens

A. Regular transcription of the reporter gene

B. One fusion protein only (Gal4-BD + Bait) - no transcription

C. One fusion protein only (Gal4-AD + Prey) - no transcription

D. Two fusion proteins with interacting Bait and Prey

Used to elucidate Protein-Protein Interactions
Challenges

- In-vitro vs in-vivo reactions

- Noise

- Noise!
Gene Regulation Simplified
Human Protein-Protein Interaction Network, via Y2H screens (Stelzl et al 2005)