CSE 527: Computational Biology
http://www.cs.washington.edu/~527

Lecture 1: Overview & Bio Review

Autumn 2005
Larry Ruzzo

He who asks is a fool for five minutes, but he who does not ask remains a fool forever.

-- Chinese Proverb

Related Courses

- Genome 540/541 (Winter/Spring)
- Stat/Biostat 578 (A 2005, TuTh 9:00-10:20)
  - Statistical Analysis of Microarrays, J. Storey
- MEDED 534, Autumn 2005
  - Biology and Informatics
- CSE590C (AWS)
  - Reading & Research in Comp. Bio.
  - Monday’s, 3:30 (MGH 284 this quarter)
  - http://www.cs.washington.edu/590c
- Combi Seminar (Genome 521; AWS)
  - Wednesday’s 1:30 K069 (sometimes 3:30 Hitch 132)
Today’s Combi

- Dr. Hamid Bolouri
- Pointillist: an open source tool for high throughput data integration
- Wednesday, September 28
- 1:30-2:30
- HSB K-069

Homework #1

- Find & read a good primer on “bio for cs” (or vice versa, as appropriate)
e.g., see ones listed on 590c page
- Post a few sentences saying
  – What you read (give me a link or citation)
  – Critique it for your meeting your needs
  – Who would it have been good for, if not you
- See class web for more details

Source:
http://www.intel.com/research/silicon/mooreslaw.htm

Source:
What’s all the fuss?

- The human genome is “finished”…
- Even if it were, that’s only the beginning
- Explosive growth in biological data is revolutionizing biology & medicine

“All pre-genomic lab techniques are obsolete”
(and computation and mathematics are crucial to post-genomic analysis)

A VERY Quick Intro To Molecular Biology

The Genome

- The hereditary info present in every cell
- DNA molecule -- a long sequence of nucleotides (A, C, T, G)
- Human genome -- about $3 \times 10^9$ nucleotides
- The genome project -- extract & interpret genomic information, apply to genetics of disease, better understand evolution, …
DNA

- Discovered 1869
- Role as carrier of genetic information - much later
- The Double Helix - Watson & Crick 1953
- Complementarity
  - A $\leftrightarrow$ T  C $\leftrightarrow$ G

Genetics - the study of heredity

- A gene -- classically, an abstract heritable attribute existing in variant forms (alleles)
- Genotype vs phenotype
- Mendel
  - Each individual two copies of each gene
  - Each parent contributes one (randomly)
  - Independent assortment

Cells

- Chemicals inside a sac - a fatty layer called the plasma membrane
- Prokaryotes (e.g., bacteria) - little recognizable substructure
- Eukaryotes (all multicellular organisms, and many single celled ones, like yeast) -
genetic material in nucleus, other organelles for other specialized functions

Chromosomes

- 1 pair of DNA molecules (+ protein wrapper)
- Most prokaryotes have just 1 chromosome
- Eukaryotes - all cells have same number of chromosomes, e.g. fruit flies 8, humans & bats 46, rhinoceros 84, …
Mitosis/Meiosis

- Most “higher” eukaryotes are diploid - have homologous pairs of chromosomes, one maternal, other paternal (exception: sex chromosomes)
- Mitosis - cell division, duplicate each chromosome, 1 copy to each daughter cell
- Meiosis - 2 divisions form 4 haploid gametes (egg/sperm)
  - Recombination/crossover -- exchange maternal/paternal segments

Proteins

- Chain of amino acids, of 20 kinds
- Proteins are the major functional elements in cells
  - Structural
  - Enzymes (catalyze chemical reactions)
  - Receptors (for hormones, other signaling molecules, odorants, …)
  - Transcription factors
  - …
- 3-D Structure is crucial: the protein folding problem

The “Central Dogma”

- Genes encode proteins
- DNA transcribed into messenger RNA
- RNA translated into proteins
- Triplet code (codons)

The Genetic Code

```
<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Pro</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
<td>Arg</td>
<td>Val</td>
</tr>
<tr>
<td>A</td>
<td>Ile</td>
<td>Phe</td>
<td>His</td>
<td>Leu</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
<td>Arg</td>
<td>Gly</td>
</tr>
</tbody>
</table>
```
**Translation: mRNA → Protein**

![Translation Diagram](image)

**Ribosomes**

![Ribosomes Diagram](image)

**Gene Structure**

- Transcribed 5’ to 3’
- Promoter region and transcription factor binding sites precede 5’
- Transcribed region includes 5’ and 3’ untranslated regions
- In eukaryotes, most genes also include introns, spliced out before export from nucleus, hence before translation

**Genome Sizes**

<table>
<thead>
<tr>
<th>Organism</th>
<th>Base Pairs</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma genitalium</td>
<td>580,073</td>
<td>483</td>
</tr>
<tr>
<td>MimiVirus</td>
<td>1,200,000</td>
<td>1,260</td>
</tr>
<tr>
<td>E. coli</td>
<td>4,639,221</td>
<td>4,290</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>12,495,682</td>
<td>5,726</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>95.5 x 10⁶</td>
<td>19,820</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>115,409,949</td>
<td>25,498</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>122,653,977</td>
<td>13,472</td>
</tr>
<tr>
<td>Humans</td>
<td>3.3 x 10⁹</td>
<td>~25,000</td>
</tr>
</tbody>
</table>
Genome Surprises

- Humans have < 1/3 as many genes as expected
- But perhaps more proteins than expected, due to alternative splicing
- There are unexpectedly many non-coding RNAs
- Many other non-coding regions are highly conserved, e.g., across all mammals

... and much more ...

- Read one of the many intro surveys or books for much more info.