A zero-knowledge based introduction to biology

Lecture 1 – Jan 6, 2015
CSE 427 Computational Biology
Instructor: Su-In Lee
TAs: Scott Lundberg, Safiye Celik
TTh 12:00-1:20 @ MGH 238

Cells: Building Block of Life
Cell, nucleus, cytoplasm, mitochondrion

- Eukaryotes:
  - Plants, animals, humans
  - DNA resides in the nucleus
  - Contain other compartments for other specialized functions

- Prokaryotes:
  - Bacteria
  - Do not contain compartments
  - Little recognizable substructure

DNA: “Blueprints” for a cell
- Genetic information encoded in long strings of double-stranded DNA (Deoxyribo Nucleic Acid)
- DNA comes in only four flavors: Adenine, Cytosine, Guanine, Thymine
  - In human, DNA is a 3 billion-long string of As, Cs, Gs and Ts
- DNA acts as the “brain” of the cell, telling the cell how to properly grow and work

Each individual has a slightly different version of the DNA sequence

![DNA Structure](image)
Nucleotide

Nucleotide, base, A, C, G, T, 3', 5'

Let's write “AGACC”!

“AGACC” (DNA)

DNA is double stranded

Strand, reverse complement

5’ AGACC 3’

3’ TCTGG 5’

DNA is always written 5’ to 3’

AGACC or GGTCT
DNA packaging

Histone, nucleosome, chromatin, chromosome, centromere, telomere

The genome

- The genome is the full set of hereditary information for an organism

- Humans have two copies of the genome into 46 chromosomes in every cell
  - $2 \times (\text{chr} 1-22 + \text{chr X/Y})$

Building an organism

Every cell has the same sequence of DNA

Subsets of the DNA sequence determine the identity and function of different cells

From DNA to organism

Proteins do most of the work in biology, and are encoded by subsequences of DNA, known as genes.
Genes & proteins

gene, transcription, translation, protein

Double-stranded DNA

5' TAGGATCGACTATATGGGATTACAAAGCATTTAGGGA...TCACCCTCTCTAGACTAGCATCTATATAAAACAGAA 3'
3' GCTATGCTGATATACCCTAATGTTTCGTAAATCCCT...GATGGGAGAGATCTGATCGTAGATATATTTTGTCTT 5'

Single-stranded RNA

translation

protein

Gene transcription

promoter

5' CATTACA... 3'
3' CTAAGT... 5'

Gene transcription

transcription factor, binding site, RNA polymerase

5' CATTACA... 3'
3' CTAGT... 5'

- **Transcription factors**: a type of protein that binds to DNA and helps initiate gene transcription.
- **Transcription factor binding sites**: short sequences of DNA (6-20 bp) recognized and bound by TFs.
- RNA polymerase binds a complex of TFs in the promoter.

Gene transcription

The two strands are separated
Gene transcription

An RNA copy of the 5'→3' sequence is created from the 3'→5' template.

RNA processing

5' cap, polyadenylation, exon, intron, splicing, UTR, mRNA

Gene structure

Transcription:  http://www.youtube.com/watch?v=DA2t5N72mgw
How many? (human genome)

- Genes: ~ 20,000
- Exons per gene: ~ 8 on average (max: 148)
- Nucleotides per exon: 170 on average (max: 12k)
- Nucleotides per intron: 5,500 on average (max: 500k)
- Nucleotides per gene: 45k on average (max: 2,2M)

From RNA to Protein

- Proteins are long strings of amino acids joined by peptide bonds
- Translation from RNA sequence to amino acid sequence performed by ribosomes
- 20 amino acids → 3 RNA letters required to specify a single amino acid

Amino acid

- There are 20 standard amino acids

Proteins

- N-terminus, C-terminus
- From 5’ → 3’ mRNA
The ribosome (a complex of protein and RNA) synthesizes a protein by reading the mRNA in triplets (codons). Each codon is translated to an amino acid.
Transcription and translation

Translation: http://www.youtube.com/watch?v=WkI_Vbwn14g&feature=related

Errors?
- What if the transcription / translation machinery makes mistakes?
- What is the effect of mutations in coding regions?

Reading frames

G C U U G U U A C G A A U U A G
G C U G U U A C G A A U U A G
G C U G U U A C G A A U U A G
G C U U G U U A C G A A U U A G

Synonymous mutation

G C U U G U U A C G A A U U A G
G C U U G U U A C G A A U U A
G C U U G U U A C G A A U U A
G C U U G U U A C G A A U U A

Ala  Cys  Leu  Arg  Ile
Ala  Cys  Leu  Arg  Ile
Missense mutation

<table>
<thead>
<tr>
<th>G</th>
<th>C</th>
<th>U</th>
<th>U</th>
<th>G</th>
<th>U</th>
<th>U</th>
<th>U</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>U</th>
<th>U</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>Cys</td>
<td>Leu</td>
<td>Arg</td>
<td>Ile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Nonsense mutation

| G | C | U | U | G | U | U | A | C | G | A | A | U | U | A | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala| Cys| Leu| Arg| Ile|

Frameshift

| G | C | U | U | G | U | U | A | C | G | A | A | U | U | A | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala| Cys| Leu| Arg| Ile|

Gene expression regulation

- When should each gene be expressed?
- **Regulate** gene expression
  
  Examples:
  - Make more of gene A when substance X is present
  - Stop making gene B once you have enough
  - Make genes C₁, C₂, C₃ simultaneously

- Why? Every cell has **same DNA** but each cell expresses **different proteins**.
- **Signal transduction**: One signal converted to another
- Cascade has “master regulators” turning on many proteins, which in turn each turn on many proteins, ...
Gene regulation

- Gene expression is controlled at many levels
  - DNA chromatin structure
  - Transcription
    - Post-transcriptional modification
    - RNA transport
    - Translation
    - mRNA degradation
    - Post-translational modification

Transcription regulation

- Much gene regulation occurs at the level of transcription.
- Primary players:
  - Binding sites (BS) in cis-regulatory modules (CRMs)
  - Transcription factor (TF) proteins
  - RNA polymerase II
- Primary mechanism:
  - TFs link to BSs
  - Complex of TFs forms
  - Complex assists or inhibits formation of the RNA polymerase II machinery

Transcription factor binding sites

- Short, degenerate DNA sequences recognized by particular TFs
- For complex organisms, cooperative binding of multiple TFs required to initiate transcription

Summary

- All hereditary information encoded in double-stranded DNA
- Each cell in an organism has same DNA
- DNA → RNA → protein
- Proteins have many diverse roles in cell
- Gene regulation diversifies protein products within different cells