After this lecture, you can...

... list the mechanisms of DNA mutation

... sort different biological information levels by conservation

... discuss the multiple roles of protein domains and draw the parallel with computer scripting

... use conservation to identify functional importance

... make and interpret consensus sequences

... make and interpret sequence logos & information content

... recognize and explain why TFBS are often palindromic

... explain the functional importance of unexpected variation
Evolution

• Replication (copy)

• Mutation (modify)

• Selection (fitness)

Mutations

• Nucleotide substitutions
  – Replication error
  – Physical or chemical reaction

• Insertions or deletions (indels)
  – Unequal crossing over during meiosis
  – Replication slippage

• Inversions or rearrangements

• Duplications of:
  – Partial or whole gene
  – Partial (polysomy) or whole chromosome (aneuploidy, polysomy)
  – Whole genome (polyploidy)

• Horizontal gene transfer (HGT)
  – Conjugation (direct transfer between Bacteria)
  – Transformation by naturally competent Bacteria
  – Transduction by bacteriophages
Phenotypic/genotypic similarity

- We exploit similarity to...
  ...identify homology (shared ancestry)
  ...determine evolutionary relationships
  ...transfer functional information
- Sequences (genotype) rarely converge
- Functions (phenotype) can converge

- Analogy
  - Similar function
- Homology
  - Similar ancestry

<table>
<thead>
<tr>
<th>Analogueous Structures (Streamline Appendages)</th>
<th>Homologous Structures (Pentadactyl Limbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor (pole)</td>
<td>Horse (trotter)</td>
</tr>
<tr>
<td>Penguin (bird)</td>
<td>Salamander (tame)</td>
</tr>
<tr>
<td>Fish</td>
<td>Human (arm and leg)</td>
</tr>
<tr>
<td>Wing</td>
<td>Cat (paw)</td>
</tr>
<tr>
<td>Flipper</td>
<td>Whale (flipper)</td>
</tr>
<tr>
<td></td>
<td>Seal (flipper)</td>
</tr>
</tbody>
</table>

Low-complexity regions

- Low-complexity regions occur in DNA and protein sequences
  - In DNA they can be cause and effect of recombination errors
  - In proteins they may be functional
  - In some cases they are used to allow for rapid evolution
    - For example in the shematrin protein in pearl oyster shells (see Figure)
Microsatellites

- If sequences do converge, they are mostly low-complexity regions
  - Microsatellites
  - Short tandem repeats (STRs)
- Most rapidly evolving characters on the genome
  - Used to distinguish individuals at short evolutionary distances

Because of selection, evolution is conservative
Orthologs and paralogs

Orthologs
- Dog (D)
- Cow (C)
- Macaque (Ma)
- Human (H)
- Chimp (Ch)
- Rat (R)
- Mouse (M)
- Chicken (G)
- Opossum (Op)

Paralogs
- Lactoperoxidase
  - Expression: Milk, saliva, tears, secretions
  - Bacteriostic activity
- Eosinophil peroxidase
  - Expression: Eosinophils
  - Microbicidal activity
- Myeloperoxidase
  - Expression: Neutrophils, phagocytes
  - Microbicidal activity
- Lactoperoxidase
  - Expression: Milk, saliva, tears, secretions
  - Bacteriostic activity
- Thyroid peroxidase
  - Expression: Thyroid surface and cytoplasm
  - Thyroid hormone biosynthesis

Conservation: sequence < structure < function

- **Sequence**
  - DNA sequence
  - Protein sequence
- **Structure**
  - Protein folding structure
- **Function**
  - Molecular function
  - Cellular function
  - Phenotype
    - Observable characteristic or trait of an organism (like morphology, development, biochemistry, physiology, or behavior)
Genetic code

- 64 codons translate into 20 amino acids, so protein sequences are more conserved than DNA sequences


(Non-) synonymous mutations

DNA level

<table>
<thead>
<tr>
<th>TTC</th>
<th>TTT</th>
<th>ATC</th>
<th>TCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lys</td>
<td>Lys</td>
<td>STOP</td>
<td>Arg</td>
</tr>
</tbody>
</table>

protein level

- Some mutations do not change the function
  - Synonymous substitutions in protein coding genes
  - Indels in non-coding (and non-regulatory) DNA
- Non-synonymous mutations change the protein sequence
- Due to the structure of the genetic code, mutations at the third nucleotide of a codon are often synonymous
Protein structure
- Protein sequence is less conserved than protein structure

Conservation of aspects of function
- Phenotypic function
  - Least conserved in evolution
  - Can rarely be predicted by sequence similarity
- Cellular function
- Molecular function
  - Best conserved in evolution
  - Can be quite reasonably predicted based on sequence similarity
Protein domains

- Protein functions can often be divided into sub-functional features
  - DNA binding
  - Zinc binding
  - Specific catalytic functions
  - ... etc
- These features may be performed by specific protein domains

![Diagram of protein domains and their functions]

Modular architecture

- Protein domains are often ...
  - ...encoded in different exons
  - ...discrete structural units
  - ...used for specific sub-functions
- Complex protein functions are thus built up of simpler sub-functions
- Like functions in a computer script, domains can easily be applied in many different contexts

![Diagram of modular architecture]
• Using sequence alignments you can identify regions that are conserved in evolution

• Conservation hints that a region is functionally important
  – A conserved region in a whole-genome alignment may be a gene or regulatory region
  – A conserved region in a single-gene alignment may be a protein domain or a short sequence motif

• Bioinformaticians use sequence conservation...
  …to detect functional elements in genomes
    • Genes
    • Transcription factor binding sites
  …to detect functional elements in proteins
    • Active sites
    • Protein-protein interaction sites
    • Ligand binding sites
  …to predict the effect of mutations in patients
    • Mutations in conserved sites are often detrimental to protein function
Sequence motifs

- A motif is a recurring pattern
  - Statistically enriched: occurs more often than expected
  - Shorter than protein domains
  - May have a function

- TATA box occurs in promoter of 24% of human genes

“Motif” rhymes with “beef”

Consensus sequences

- Sometimes it is handy to summarize hundreds of aligned sequences
- The consensus sequence contains the most frequent residue at each position
- Ambiguity characters can be used
  - Protein: only X (any amino acid possible, no gap)
  - DNA: all ambiguity characters

A: Adenine
C: Cytosine
G: Guanine
T: Thymine
U: Uracil

V, W, Y: pyrimidine (C, G, or T)
R, S, K: purine (A, G, or T)
M: adenosine (A or G)
N: a general agreement: there is a growing consensus that the current
DNA sequence logos

- Sequence logo has more detail than consensus seq
- At each position, possible nucleotides are shown by stacked letters
  - Letter heights relative to frequencies $p_i$ ($i = A, C, G, T$)
- The total stack height shows the conservation
  - Information content at given position (in bits):
    $$ I = \sum_{i \in \{A,C,G,T\}} p_i \log_2 \left( \frac{p_i}{1/4} \right) $$
    - Maximum information in a completely conserved position (e.g. always T)
      - $p_A = 0; p_C = 0; p_G = 0; p_T = 1$
      - Assume that $0 \log_2 (0) = 0$
      - $I = 0 + 0 + 0 + \log_2 (4) = 2$
    - Minimum information in a completely unconserved position (random)
      - $p_A = 0.25; p_C = 0.25; p_G = 0.25; p_T = 0.25$
      - $I = 0.25 \log_2 (1) + 0.25 \log_2 (1) + 0.25 \log_2 (1) + 0.25 \log_2 (1) = 0 + 0 + 0 + 0 = 0$
Protein sequence logos

- Amino acid sequence logos summarize aligned protein sequences
  - Letter heights relative to amino acid frequencies $p_i$
  - The total stack height shows the conservation
  - Information content at position $k$ (in bits):
    \[
    I = \sum_{i \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}} p_i \log_2 \left( \frac{p_i}{1/20} \right)
    \]
  - **Maximum information in a completely conserved position**
    \[
    I = (19 \cdot 0) + \log_2 (20) = 4.3219
    \]
  - **Minimum information in a completely random position**
    \[
    I = 20 \cdot (0.05 \log_2 (1)) = 0
    \]

Weblogo

- Weblogo is a webserver to create sequence logos from a multiple alignment: [weblogo.berkeley.edu](http://weblogo.berkeley.edu)
Transcription factor binding sites

- Transcription factor binding sites (TFBSs) often contain palindromic DNA sequence motifs
- Transcription factors are often homodimeric proteins, where both halves of the dimer bind to opposite strands of the helical DNA

Homodimeric transcription factor proteins

Helical palindromic DNA sequence

Exercise

\[ I = \sum_{i \in \{A,C,G,T\}} p_i \log_2 \left( \frac{p_i}{1/4} \right) \]

\[ I = \sum_{i \in \{A,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,Y,W,X\}} p_i \log_2 \left( \frac{p_i}{1/20} \right) \]

a. Which positions are fully conserved?
b. Which positions are fully random?
c. Why is the y-axis different between the two sequence logos?
d. Give the maximum stack height for DNA sequence logos (in bits).
e. Give the maximum stack height for protein sequence logos.
f. Give both the consensus sequences.
Meaningful sequence variation

• In some cases, binding sites are much less conserved than expected (positive selection, rapid evolution)
  – For example binding of a virus to its host membrane proteins

• Alignments can be used to identify such hypervariable regions or proteins
  – Discover virus-host interaction proteins

Aligning metagenomic reads from many different samples to the crAssphage genome reveals metagenomic islands:

Metagenomic islands

• Regions in a genome that are highly divergent from the sequences found in metagenomes

### Table 1. Annotation and frequency of the most frequent genes found in MHVs

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Frequency</th>
<th>Inferred Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phage tail fiber protein</td>
<td>12</td>
<td>Host recognition</td>
</tr>
<tr>
<td>T7 terminase subunit</td>
<td>10</td>
<td>DNA packaging</td>
</tr>
<tr>
<td>CRISPR/Cas domain containing protein</td>
<td>10</td>
<td>Host recognition</td>
</tr>
<tr>
<td>PyrG-dependent pyruvate superfamily protein</td>
<td>9</td>
<td>Host recognition</td>
</tr>
<tr>
<td>Phage tail fiber assembly protein</td>
<td>6</td>
<td>DNA packaging</td>
</tr>
<tr>
<td>Phage tail fiber adhesin</td>
<td>3</td>
<td>Host recognition</td>
</tr>
<tr>
<td>Phage tail fiber adhesin Omp88</td>
<td>3</td>
<td>Phage structure</td>
</tr>
<tr>
<td>DNA binding domain</td>
<td>3</td>
<td>DNA binding</td>
</tr>
<tr>
<td>Phage internal content protein</td>
<td>3</td>
<td>Host cell penetration</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>252</td>
<td>Unknown</td>
</tr>
<tr>
<td>Others</td>
<td>46</td>
<td>Other</td>
</tr>
<tr>
<td>Total</td>
<td>380</td>
<td></td>
</tr>
</tbody>
</table>

(a) Schematic drawing of T2 bacteriophage
(b) An electron micrograph of T2 bacteriophage infecting E. coli
HIV-1

- HIV infects helper T cells of the immune system
- Loss of immune cells impairs immune responses, leading to acquired immune deficiency syndrome (AIDS)
- HIV eludes the immune system by mutating very rapidly

Timeline of HIV-1 infection

- Primary Infection
- Acute HIV syndrome
- Wide dissemination of virus
- Seeding of lymphoid organs
- Clinical Latency
- Opportunistic Diseases
- Constitutional Symptoms
- Death

CD4 T Lymphocyte Count (cells/mm³)

0 100 200 300 400 500 600 700 800 900 1000 1100 1200
0 10^3 10^4 10^5 10^6 10^7

Weeks Years

HIV RNA Copies per ml Plasma
Immune escape

Weeks after infection

<table>
<thead>
<tr>
<th>Antibodies to variant 1 appear</th>
<th>Antibodies to variant 2 appear</th>
<th>Antibodies to variant 3 appear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variant 1</td>
<td>Variant 2</td>
<td>Variant 3</td>
</tr>
</tbody>
</table>

Amount of virus in blood

Weeks after infection

25 26 27 28

0 0.5 1.0 1.5

Which HIV-1 proteins should we use in vaccine?

Key to Terms

- **HIV capsid**: HIV's bullet-shaped core that contains HIV RNA
- **HIV envelope**: Outer surface of HIV
- **HIV enzymes**: Proteins that carry out steps in the HIV life cycle
- **HIV glycoproteins**: Protein “spikes” embedded in the HIV envelope
- **HIV RNA**: HIV’s genetic material
Which HIV-1 proteins should we use in vaccine?

- HIV Envelope
- HIV Capsid

Werkcolleges

- Sla over: vragen 15-18 en 26