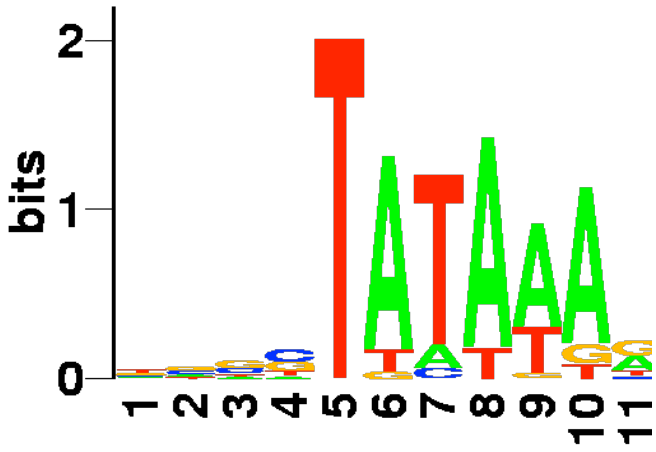
Sequence conservation



Bas E. Dutilh
Systems Biology: Bioinformatic Data Analysis
Utrecht University, February 16th 2017

Utrecht Bioinformatics Center

Universiteit Utrecht

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

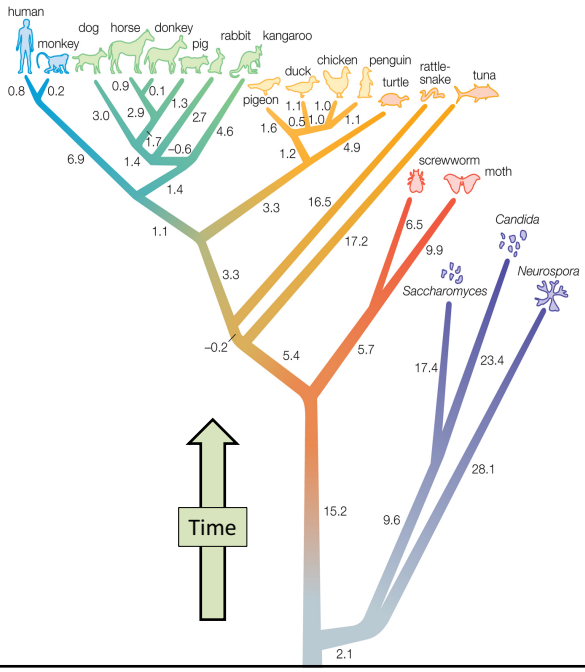
… explain the functional importance of unexpected variation

Evolution

- Replication (copy)

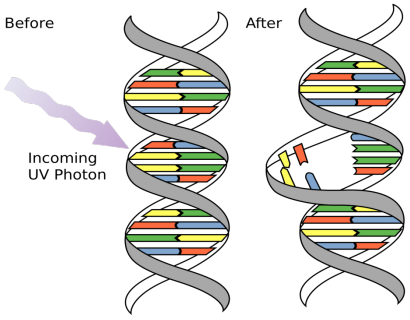
- Mutation (modify)

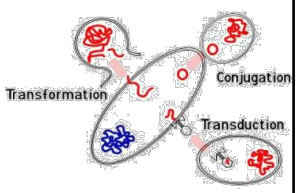
- Selection (fitness)



Mutations

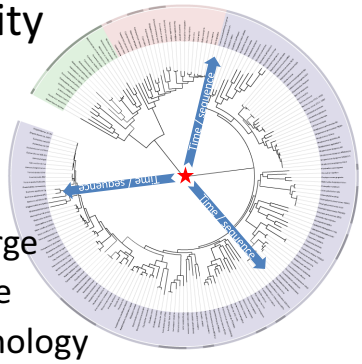
- Nucleotide substitutions
  - Replication error
  - Physical or chemical reaction
- Insertions or deletions (indels)
  - Unequal crossing over during meiosis
  - Replication slippage
- Inversions or rearrangements
- Duplications of:
  - Partial or whole gene
  - Partial (polysomy) or whole chromosome (aneuploidy, polysomy)
  - Whole genome (polyploidy)
- Horizontal gene transfer (HGT)
  - Conjugation (direct transfer between Bacteria)
  - Transformation by naturally competent Bacteria
  - Transduction by bacteriophages

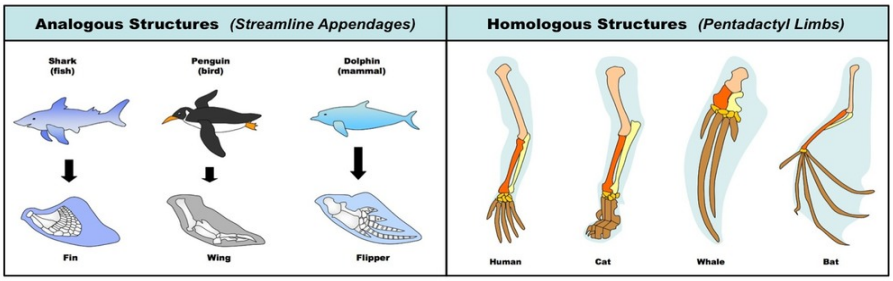




Phenotypic/genotypic similarity

- We exploit similarity to…
  - …identify homology (shared ancestry)
  - …determine evolutionary relationships
  - …transfer functional information
- Sequences (genotype) rarely converge
- Functions (phenotype) can converge
- Analogy
  - Similar <u>function</u>
- Homology
  - Similar <u>ancestry</u>

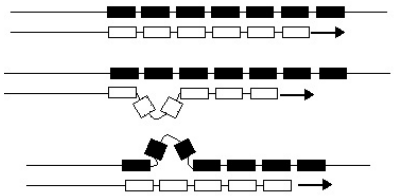




Low-complexity regions

- <u>Low-complexity regions</u> 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)

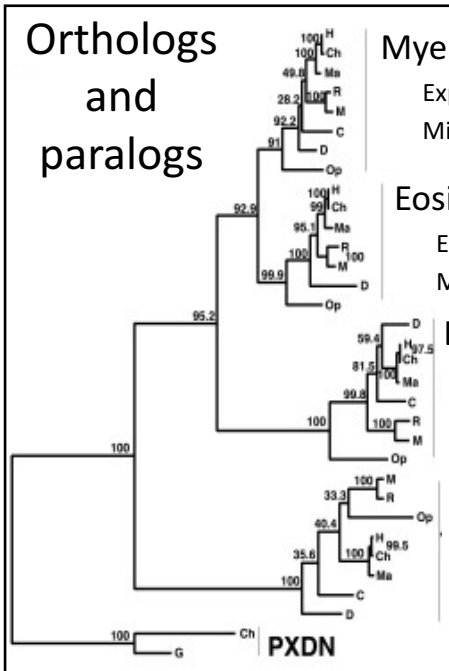
Microsatellites



- 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



Myeloperoxidase

Expression: Neutrophils, phagocytes

Microbicidal activity

Eosinophil peroxidase

Expression: Eosinophils

Microbicidal activity

Lactoperoxidase

Expression: Milk, saliva, tears, secretions

Bacteriosttic activity

Thyroid peroxidase

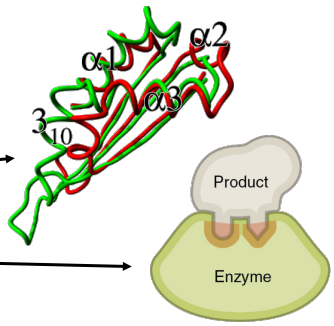
Expression: Thyroid surface and cytoplasm

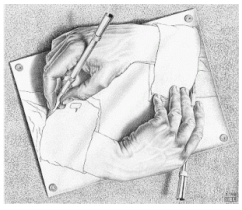
Thyroid hormone biosynthesis

Dog (D); Cow (C); Macaque (Ma); Human (H); Chimp (Ch); Rat (R); Mouse (M), Chicken (G), and Opossum (Op).

Conservation: sequence < structure < function

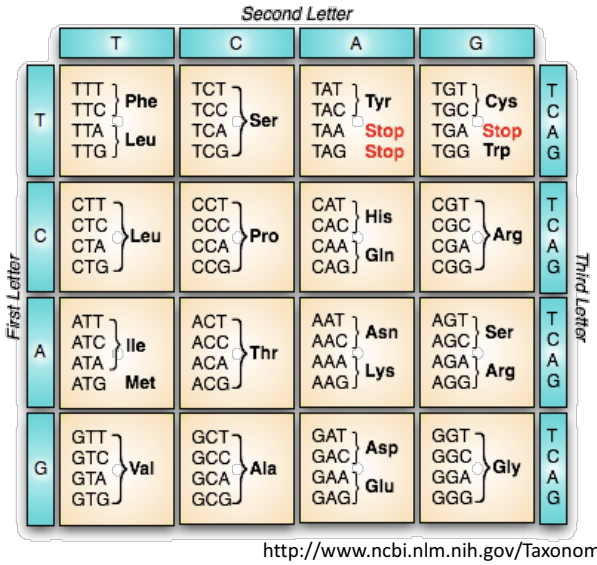
- Sequence
  - DNA sequence
  - Protein sequence
- Structure
  - Protein folding structure
- Function
  - Molecular function
  - Cellular function
  - Phenotype
    - = Observable characteristic or trait of an organism (like morphology, development, biochemistry, physiology, or behavior)





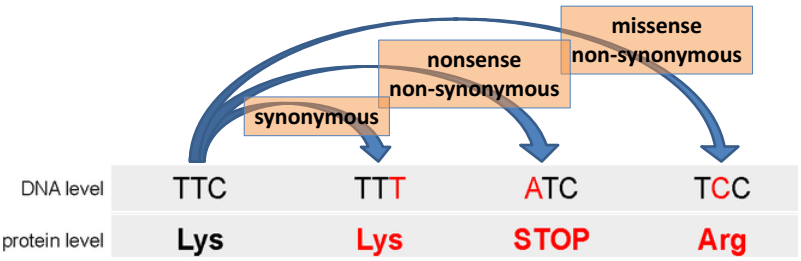
Genetic code

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



http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

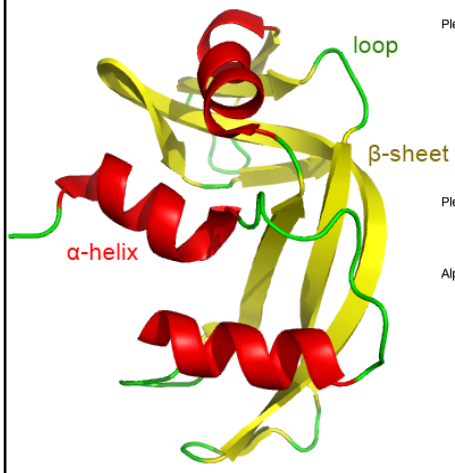
(Non-) synonymous mutations

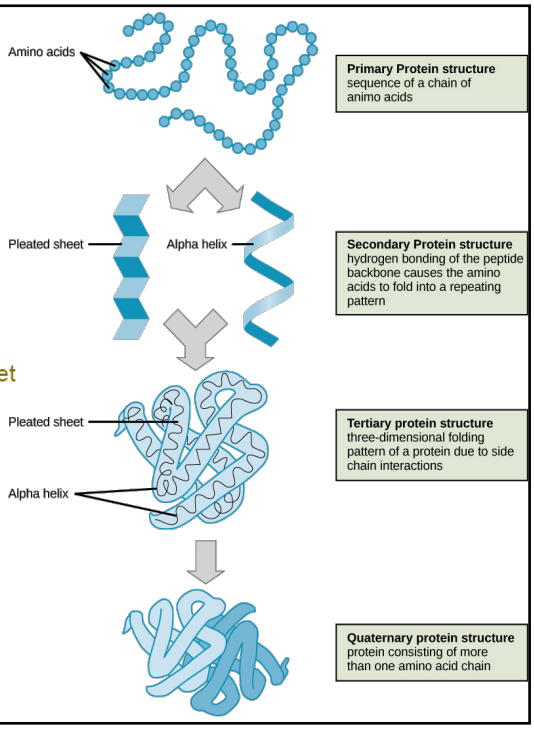


- Some mutations do not change the function
  - Synonymous substitutions in protein coding genes
  - Indels in non-coding (and non-regulatory) DNA
- Non-synonymous mutations change the protein sequence
- Due to the structure of the genetic code, mutations at the third nucleotide of a codon are often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

Protein structure

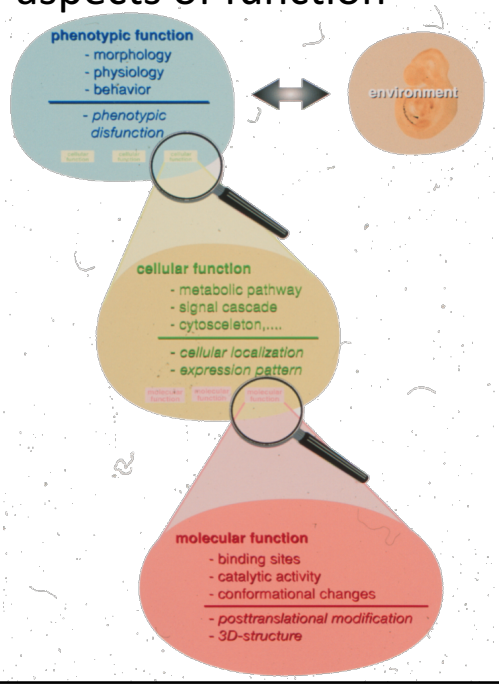
- Protein sequence is less conserved than protein structure





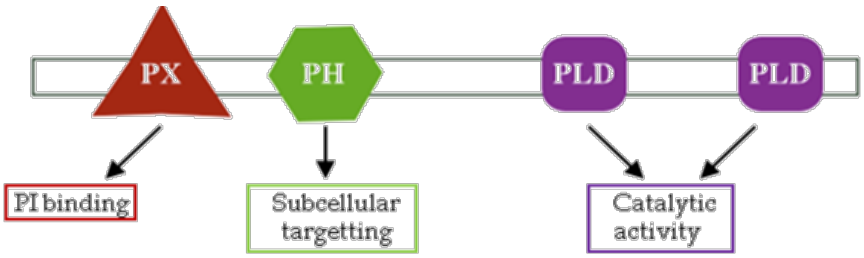
Conservation of aspects of function

- 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



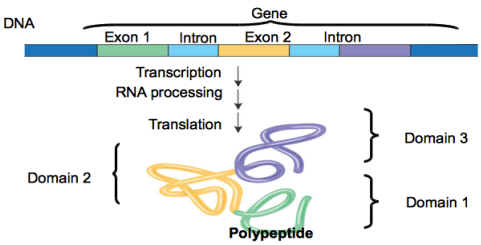
Protein domains

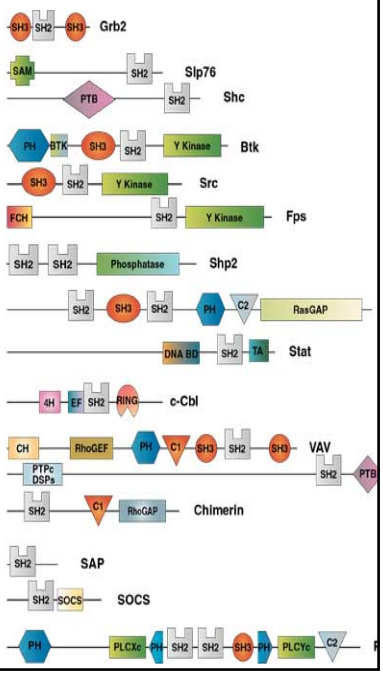
- Protein functions can often be divided into sub-functional features
  - DNA binding
  - Zinc binding
  - Specific catalytic functions
  - … etc
- These features may be performed by specific protein domains



Modular architecture

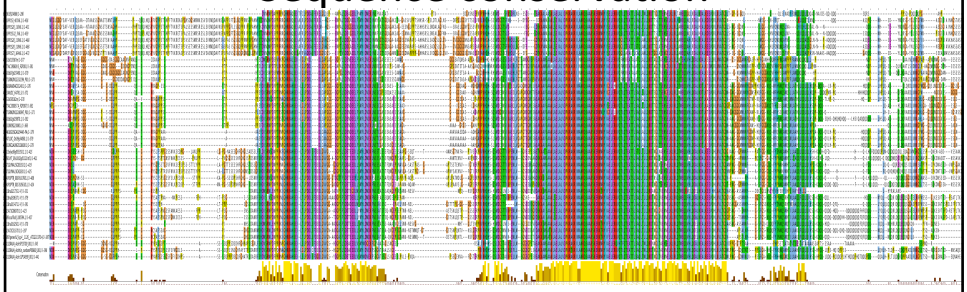
- Protein domains are often …
  - …encoded in different exons
  - …discrete structural units
  - …used for specific sub-functions





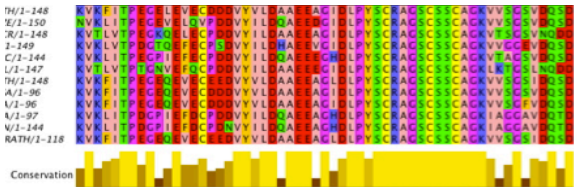
- Complex protein functions are thus built up of simpler sub-functions
- Like functions in a computer script, domains can easily be applied in many different contexts

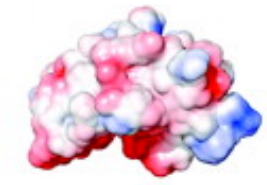
Sequence conservation

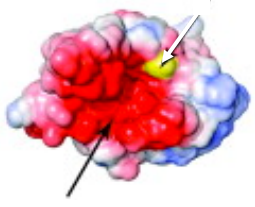


- Using sequence alignments you can identify regions that are conserved in evolution
- Conservation hints that a region is functionally important
  - A conserved region in a whole-genome alignment may be a gene or regulatory region
  - A conserved region in a single-gene alignment may be a protein domain or a short sequence motif

Conserved is probably important







Binding site

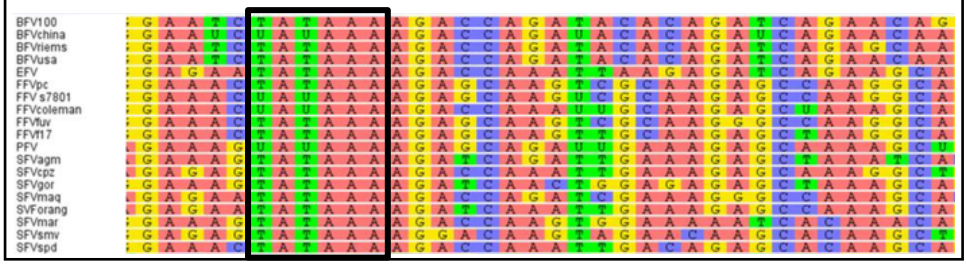
- Bioinformaticians use sequence conservation…
  - …to detect functional elements in genomes
    - Genes
    - Transcription factor binding sites
  - …to detect functional elements in proteins
    - Active sites
    - Protein-protein interaction sites
    - Ligand binding sites
  - …to predict the effect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

Low High

Sequence motifs

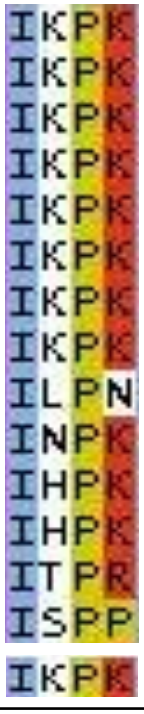
- A motif is a recurring pattern
  - Statistically enriched: occurs more often than expected
  - Shorter than protein domains
  - May have a function
- TATA box occurs in promoter of 24% of human genes

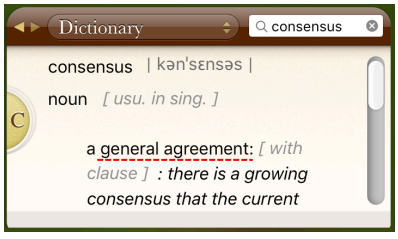
"Motif" rhymes with *"beef"*



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



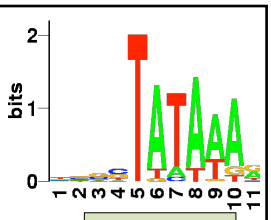


A	Adenine	
G	Guanine	
C	Cytosine	
T	Thymine	
U	Uracil	
N	unknown	(A or C or G or T)
Y	pYrimidine	(C or T)
R	puRine	(A or G)
W	"Weak"	(A or T)
S	"Strong"	(C or G)
K	"Keto"	(T or G)
M	"aMino"	(C or A)
B	not A	(C or G or T)
D	not C	(A or G or T)
H	not G	(A or C or T)
V	not T	(A or C or G)

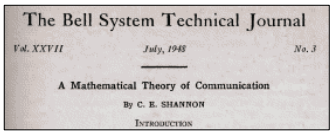
DNA sequence logos

- 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):

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



The TATA box


The Bell System Technical Journal
Vol. XXVII July, 1948 No. 3
A Mathematical Theory of Communication
By C. E. SHANNON
Introduction

- 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)
  - $p_A = 0.25$; $p_C = 0.25$; $p_G = 0.25$; $p_T = 0.25$

$$I = 0.25 \log_2(1) + 0.25 \log_2(1) + 0.25 \log_2(1) + 0.25 \log_2(1)$$

$$I = 0 + 0 + 0 + 0 = \mathbf{0}$$

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):

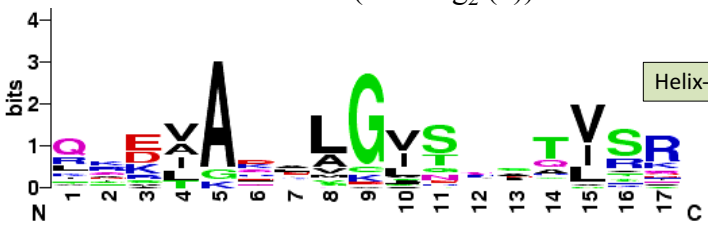
$$I = \sum_{i \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}} p_i log_2(\frac{p_i}{1/20})$$

- Maximum information in a completely conserved position

$$I = (19 \cdot 0) + \log_2(20) = \mathbf{4.3219}$$

- Minimum information in a completely random position

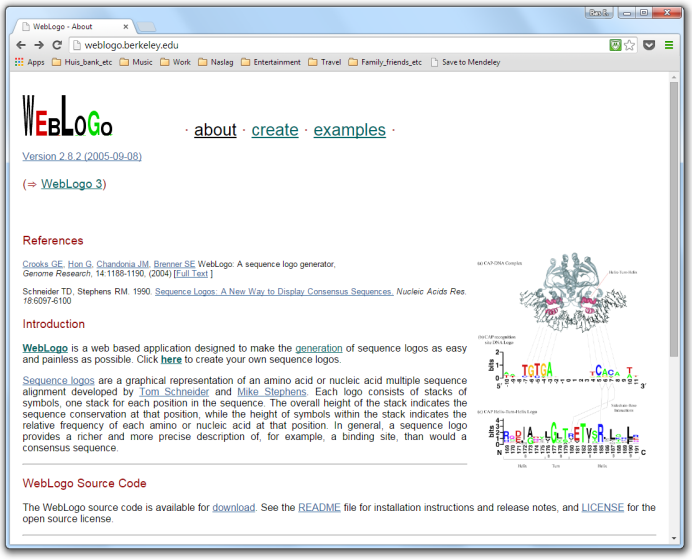
$$I = 20 \cdot (0.05 \log_2(1)) = \mathbf{0}$$



Helix-turn-helix motifs

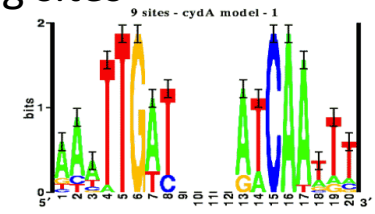
Weblogo

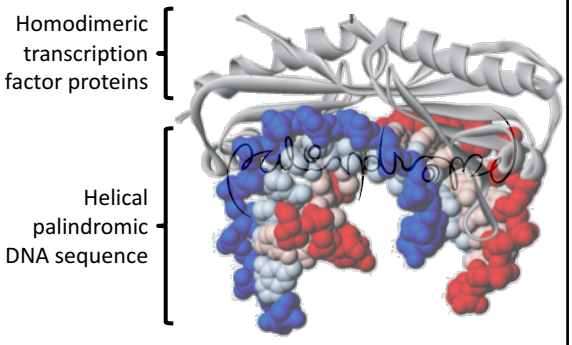
- Weblogo is a webserver to create sequence logos from a multiple alignment: weblogo.berkeley.edu



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

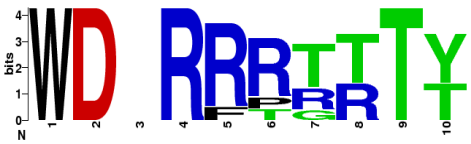




Exercise

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

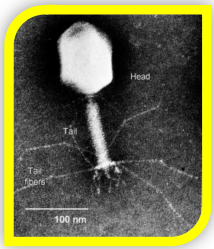
$$I = \sum_{i \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}} p_i log_2(\frac{p_i}{1/20})$$



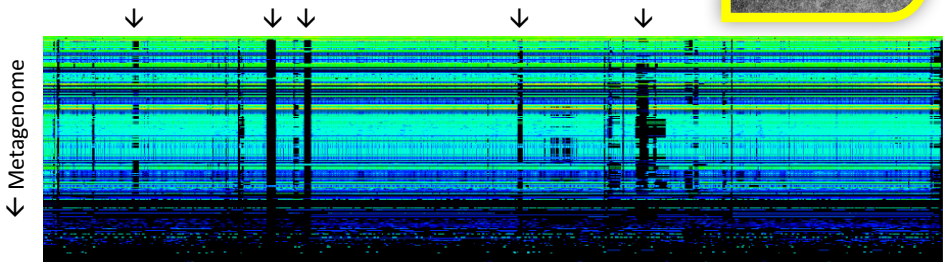
a. Which positions are fully conserved?
b. Which positions are fully random?
c. Why is the y-axis different between the two sequence logos?
d. Give the maximum stack height for DNA sequence logos (in bits).
e. Give the maximum stack height for protein sequence logos.
f. Give both the consensus sequences.

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:

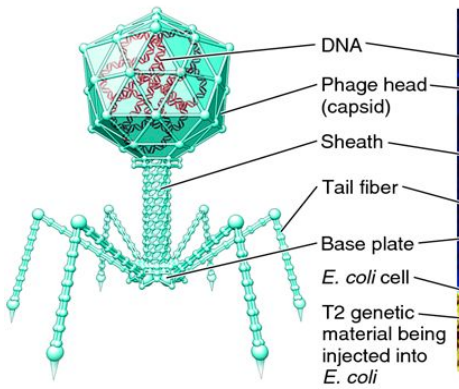


Metagenomic islands

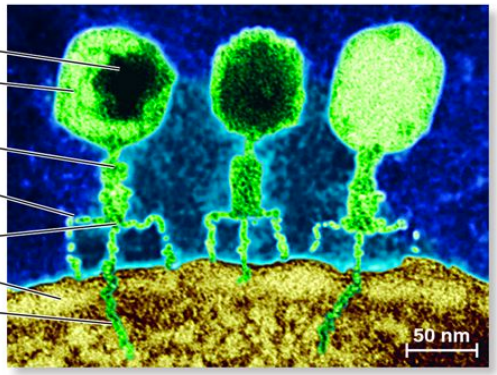
- Regions in a genome that are highly divergent from the sequences found in metagenomes

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

Annotation	Frequency	Inferred function
Putative carbohydrate binding domain containing protein	14	Host recognition
Tail fiber protein	12	Host recognition
Terminase small subunit	10	DNA packaging
C1q like domain containing protein	10	Host recognition
Lectin-like domain containing protein	9	Host recognition
Fe(II)-dependent oxygenase superfamily protein	8	Host recognition
Terminase large subunit	6	DNA packaging
Phage-like element PbsX protein XkdW	4	unknown
Putative tail fiber assembly protein	4	Phage structure
Fibrinogen-like coiled coil protein	3	Host recognition
Phage tail fibre adhesin Gp38	3	Host recognition
Tape measure protein	3	Phage structure
DNA binding domain	3	DNA binding
Putative internal virion protein	3	Host cell penetration
Hypothetical protein	252	unknown
Others	46	other
total	390	

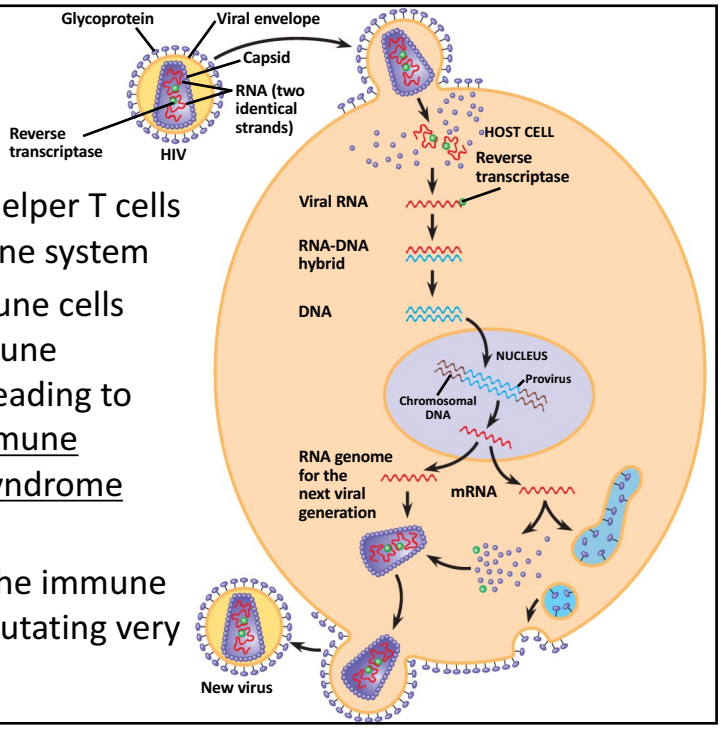


(a) Schematic drawing of T2 bacteriophage



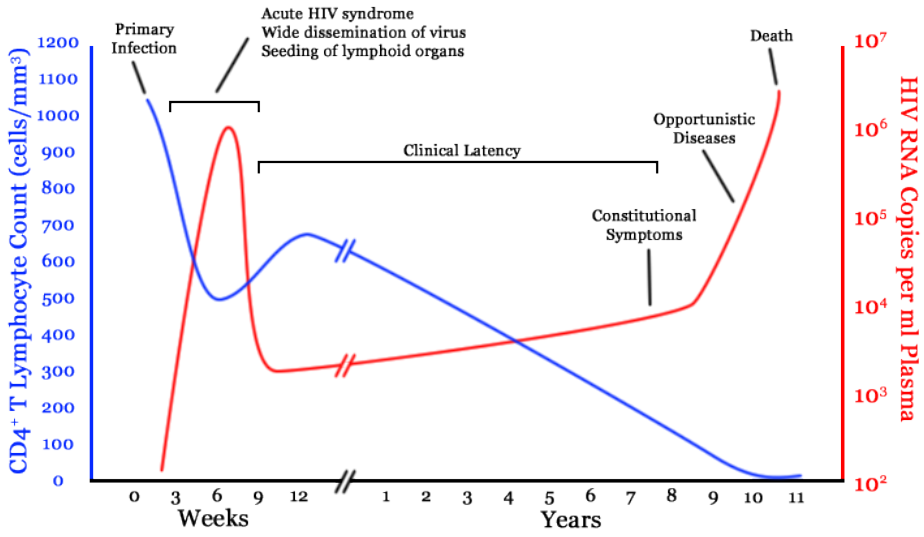
(b) An electron micrograph of T2 bacteriophage infecting *E. coli*

HIV-1

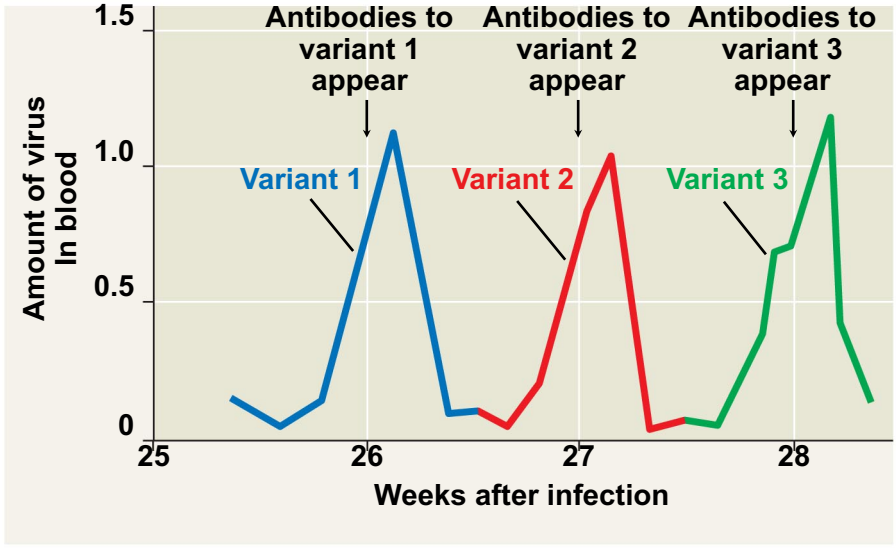


- HIV infects helper T cells of the immune system
- Loss of immune cells impairs immune responses, leading to <u>acquired immune deficiency syndrome</u> (AIDS)
- HIV eludes the immune system by mutating very rapidly

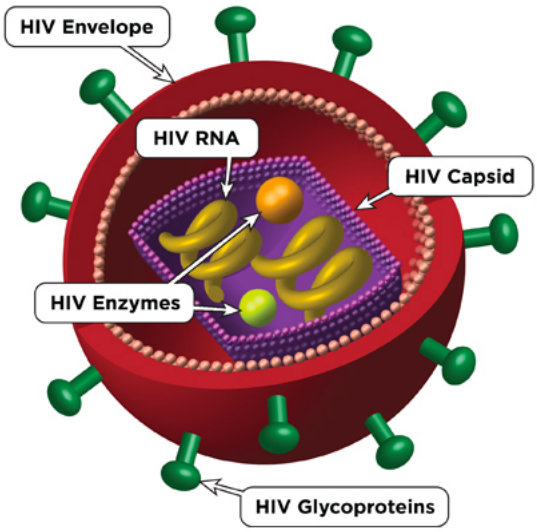
Timeline of HIV-1 infection



Immune escape



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

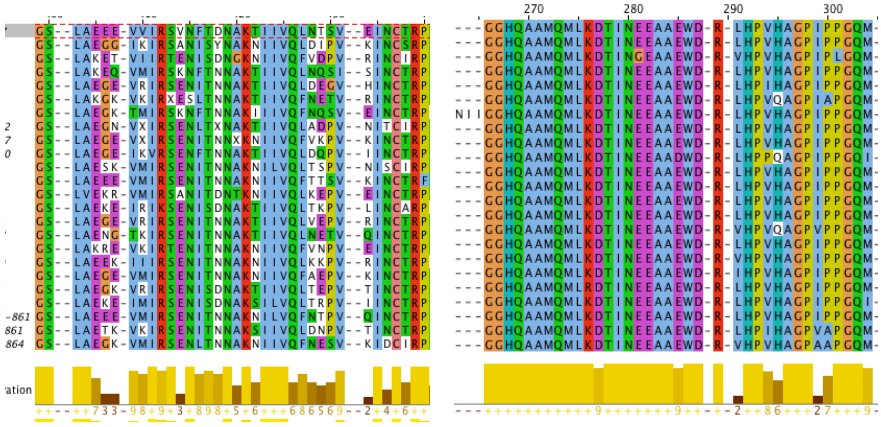
HIV enzymes: Proteins that carry out steps in the HIV life cycle

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