BIO-INFORMATICA

COURSE 3B_ VERGELIJKENDE GENOOMANALYSE

WEEK 3: CODON BIAS

HAN_UNIVERSITY OF APPLIED SCIENCES

OVERZICHT ONDERZOEKSPLAN 3B

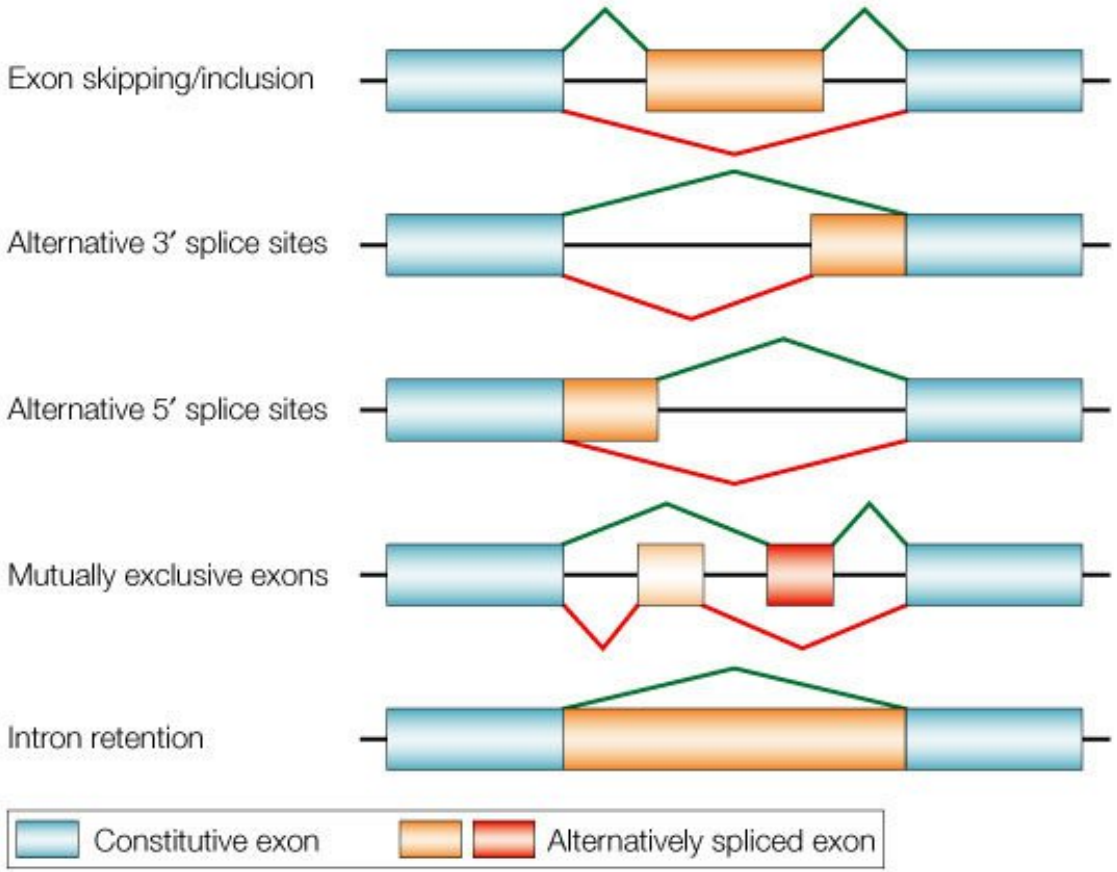
Week	Onderwerp	Activiteiten	M&M
1	Oriëntatie op onderwerp Sequenties verzamelen	Inlezen Stap 1	NCBI database
2	Genen en GC percentage	Stap 2	Python script
3	Codon gebruik	Stap 3	Python script
4	Eiwit karakterisering / aminozuren	Stap 4	Python script
5	Kenmerken oppervlakte proteïnen	Stap 5	Python script
6	Fylogenetisch onderzoek	Stap 6	Bioinf tools
7	Onderzoeksverslag		

DE WEEK

- Herhaling vorige weken
- Genen vinden bij:
  - Eukaryoten en prokaryoten.
- Codon gebruik
- Codon bias
- Evolutie

- Pevsner Hfst 17, pg. 819-820
- Pevsner Hfst 20, pg. 969 CpG islands

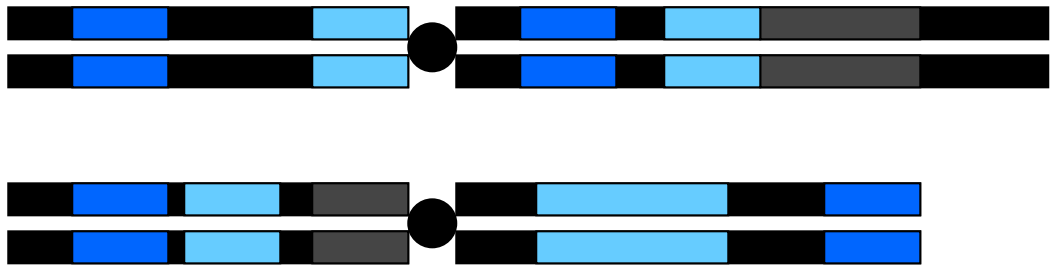
HERHALING GENEN, EXONEN EN ALTERNATIVE SPLICING



http://www.nature.com/scitable/content/a-schematic-representation-of-alternative-splicing-14263384

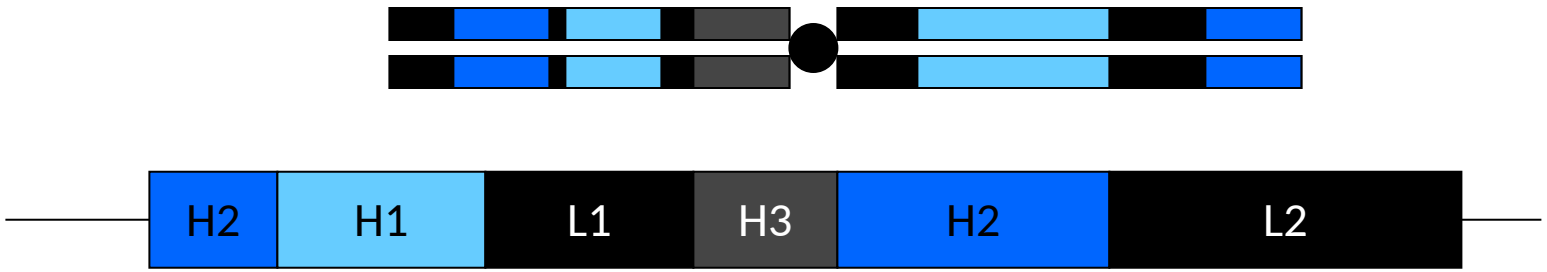
HERHALING VORIGE WEEK GC%

- GC% van de DNA sequentie varieert:
  - Tussen genomen (organismen)
  - Coderende sequenties
    - Genen hogere GC ratio
  - Op het genoom
    - Mozaïek achtige formatie: regio's op het genoom van GC-rijke en niet rijke regio's die we isochores noemen



HERHALING VORIGE WEEK ISOCHORES

Genomic core	H1, H2 and H3 (47%, 52% >52%)
“Empty” space	L1 and L2 (<40%)



HERHALING VORIGE WEEK HUMAN ISOCHORES

L1 en L2 arm in GC (~40%):

- Gen arm
- 85% van weefsel specifieke genen

H1, H2 en H3 rijk in GC (~50%):

- Gen rijk
- 80% huishoudgenen
- Promoter dichtbij transcriptie
- Kortere intronen en genen
- Andere codon bias (Meer G+C)
- Less long repeated sequences (LINEs)

CODON GEBRUIK EN BIAS

- 64 codons, 61 coderen voor aminozuren.
- Er zijn 20 aminozuren.
- Synonieme codons (triplets) in coding DNA kunnen bias vertonen in hun gebruik.
  - Kan verschillen tussen genen in een organisme en tussen organismen.
  - Voorkeur voor bepaalde codons in genen met hoge expressie.
  - Ligt aan de tRNA repertoire/beschikbare tRNA en andere factoren.
  - Mogelijk selectie voor bepaalde codons vanwege door translatie optimalisatie en vouw stabiliteit van mRNA.

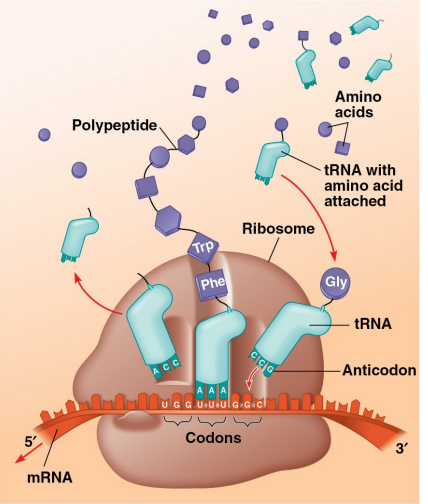


Fig. 17.15, Cambell Biology, 2021 Global edition

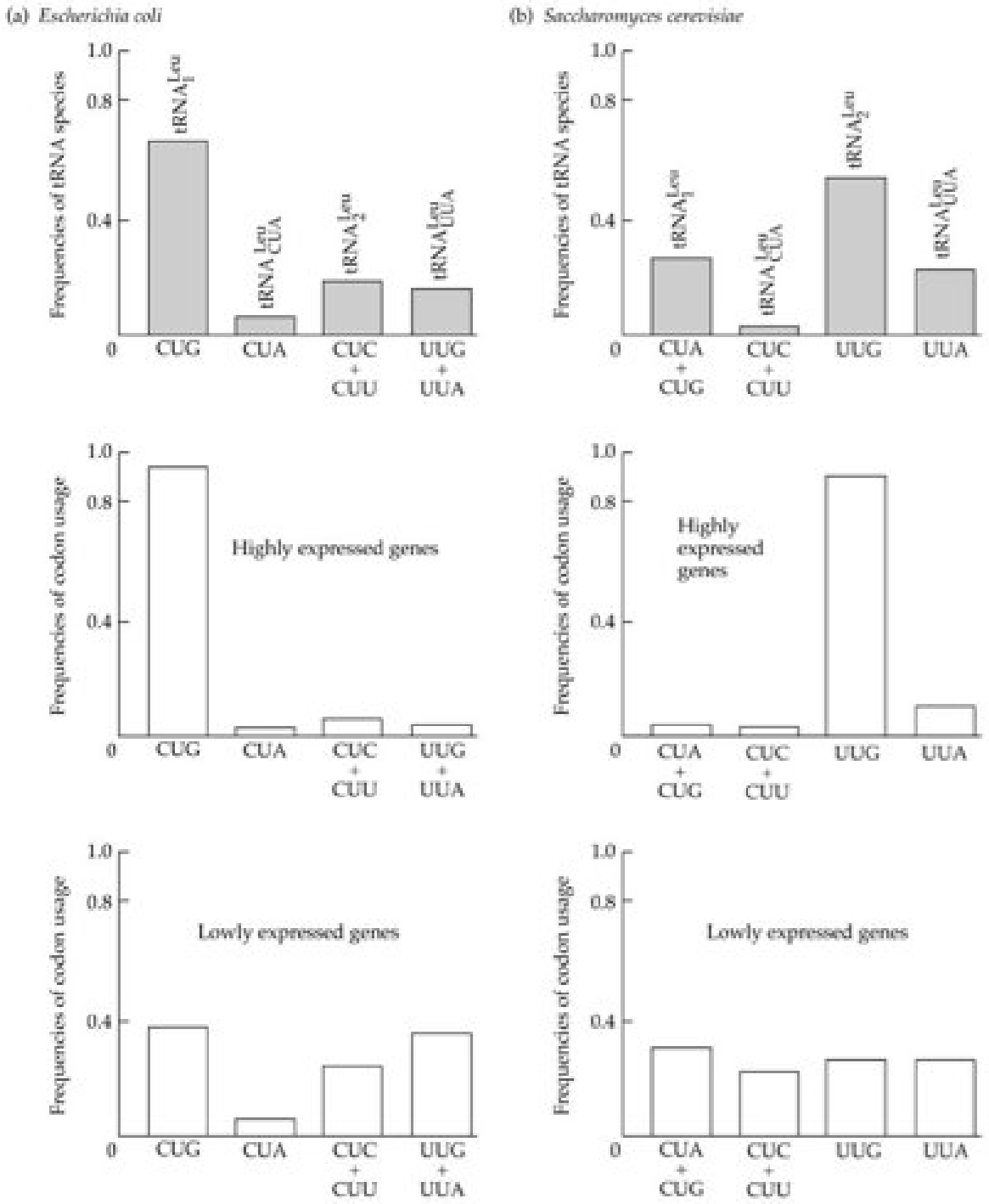
Second mRNA base

First mRNA base (5′ end of codon)	U	C	A	G	Third mRNA base (3′ end of codon)
U	UUU Phe (F)	UCU Ser (S)	UAU Tyr (Y)	UGU Cys (C)	U
	UUC Phe (F)	UCC Ser (S)	UAC Tyr (Y)	UGC Cys (C)	C
	UUA Leu (L)	UCA Ser (S)	UAA Stop	UGA Stop	A
	UUG Leu (L)	UCG Ser (S)	UAG Stop	UGG Trp (W)	G
C	CUU Leu (L)	CCU Pro (P)	CAU His (H)	CGU Arg (R)	U
	CUC Leu (L)	CCC Pro (P)	CAC His (H)	CGC Arg (R)	C
	CUA Leu (L)	CCA Pro (P)	CAA Gln (Q)	CGA Arg (R)	A
	CUG Leu (L)	CCG Pro (P)	CAG Gln (Q)	CGG Arg (R)	G
A	AUU Ile (I)	ACU Thr (T)	AAU Asn (N)	AGU Ser (S)	U
	AUC Ile (I)	ACC Thr (T)	AAC Asn (N)	AGC Ser (S)	C
	AUA Ile (I)	ACA Thr (T)	AAA Lys (K)	AGA Arg (R)	A
	AUG Met (M) or start	ACG Thr (T)	AAG Lys (K)	AGG Arg (R)	G
G	GUU Val (V)	GCU Ala (A)	GAU Asp (D)	GGU Gly (G)	U
	GUC Val (V)	GCC Ala (A)	GAC Asp (D)	GGC Gly (G)	C
	GUA Val (V)	GCA Ala (A)	GAA Glu (E)	GGA Gly (G)	A
	GUG Val (V)	GCG Ala (A)	GAG Glu (E)	GGG Gly (G)	G

Fig. 17.6, Cambell Biology, 2021 Global edition

GEN EXPRESSIE BEÏNVLO DOOR CODON GEBRUIK

- Hoe makkelijk een gen wordt getransleerd ligt aan tRNA’s die aanwezig zijn.

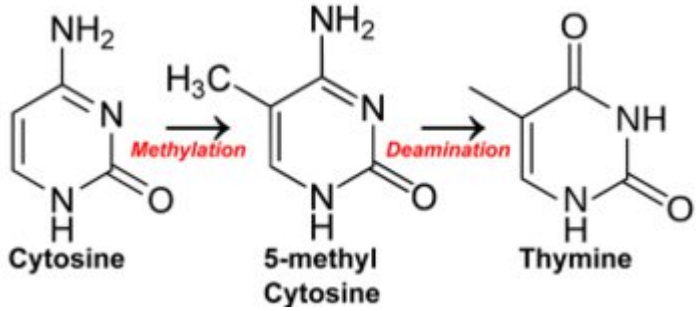


CODON GEBRUIK EN BIAS

- Codon gebruik = gebruik van de codons in een gen of organisme.
- Codon bias = de voorkeur van een organisme voor het gebruik van bepaalde codons ten opzichte van andere codons.
- Correlatie tussen
  - GC content op de derde positie van een codon in een gen GC3.
    - GC3 ~40% in L1 en L2
    - GC3 ~80% in H1, H2 en H3
  - GC content in een bepaalde regio's.

CpG ISLANDS

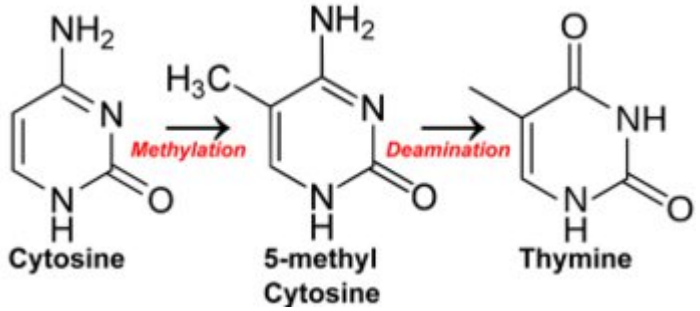
- Regulatie element: Cs gevolgd door Gs.
  - *5'—C—phosphate—G—3'*
- Met C (naast een G) vaak gemethyleerd.
- Methylering resulteert in werving van eiwit complexen die zorgen dat DNA strakker oprolt.
  - Dit remt actieve transcriptie (epigenetica).



CpG site - Wikipedia

CpG ISLANDS

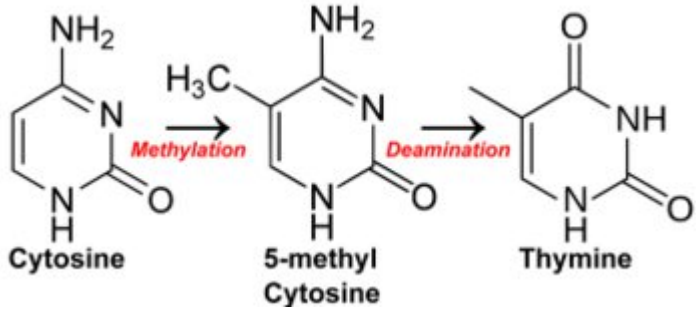
- Boven (5’) regulatie regio's bij transcriptie start sites van huishoud genen CpG met hoge dichtheid aan niet gemethyleerde nucleotiden.



CpG site - Wikipedia

CpG ISLANDS

- Gemetyleerde C’s muteren makkelijk naar T’s.
- Dit kan eilanden vormen doordat in regio’s waar methylatie niet vaak voorkomt op het genoom, CpG overblijft (C’s worden niet gemetyleerd).
- In veel zoogdieren worden CpG eilanden in associatie gebracht met de start van genen en kan dit gebruikt worden voor het voorspellen en de annotatie van genen.



CpG site - Wikipedia

CpG ISLANDS

- CpG eilanden hebben een GC content ≥ 50% en lengte van ≥ 200 bp en een ratio van > 0.6.
- In dieren zijn 70%-80% CpG's gemethyleerd.
- In de mens heeft 70% van promoteren vlakbij transcriptie start sites van een gen CpG eilanden.

(b) CpG island associated with *HBA1*

>chr16:226174-227254
CGTCCGGGTGCGCGCATTCCTCTCCGCCCCAGGATTGGGCGAAGCCTCCCGGCTCGCACT
CGCTCGCCCGTGTGTTCCCCGATCCCGCTGGAGTCGATGCGCGTCCAGCGCGTGCCAGGC
CGGGGCGGGGGTGCGGGCTGACTTTCTCCCTCGCTAGGGACGCTCCGGCGCCCGAAAGGA
AAGGGTGGCGCTGCGCTCCGGGGTGCACGAGCCGACAGCGCCCGACCCCAACGGGCCGGC
CCCGCCAGCGCCGCTACCGCCCTGCCCCCGGGCGAGCGGGATGGGCGGGAGTGGAGTGGC
GGGTGGAGGGTGGAGACGTCCTGGCCCCCGCCCCGCGTGCACCCCCAGGGGAGGCCGAGC
CCGCCGCCCCGGCCCCGCGCAGGCCCCGCCCGGGACTCCCCTGCGGTCCAGGCCGCGCCCC
GGGCTCCGCGCCAGCCAATGAGCGCCGCCCGGCCGGGCGTGCCCCCGCGCCCCAAGCATA
AACCCTGGCGCGCTCGCGGCCCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCA
CCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCG
CGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTGCTCCG
ACCCGGGCTCCTCGCCCGCCCGGACCCACAGGCCACCCTCAACCGTCCTGGCCCCGGACC
CAAACCCCACCCCTCACTCTGCTTCTCCCCGCAGGATGTTCCTGTCCTTCCCCACCACCA
AGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCA
AGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGC
TGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGG
TGAGCGGCGGGCCGGGAGCGATCTGGGTCGAGGGGCGAGATGGCGCCTTCCTCGCAGGGC
AGAGGATCACGCGGGTTGCGGGAGGTGTAGCGCAGGCGGCGGCTGCGGGCCTGGGCCCTC
G

FIGURE 8.15 CpG islands are associated with the regulation of expression of many eukaryotic genes. (a) The alpha globin gene cluster on human chromosome 16 is shown (in a window of 35,000 base pairs of chr16:200,001–235,000 on the UCSC Genome Browser). Each of the five genes has an associated CpG island, defined as having a GC content of 50% or greater, a length greater than 200 base pairs, and a ratio >0.6 of observed to expected CpG dinucleotides. (b) By clicking on the *HBA2* CpG island, its DNA sequence (chr16:222,370–223,447) is accessed. CpG dinucleotides are highlighted in pink.

Source: http://genome.ucsc.edu, courtesy of UCSC.

CODON GEBRUIK EN BIAS

- Codons met CpG worden vaak vermeden.
  - Low-usage codons
  - Bijv. Arginine (CGA en CGG)
- Er lijkt een voorkeur te zijn voor naast elkaar gelegen codon:
  - NNG en GNN

Amino Acid	Codon	*Escherichia coli*
Leucine	UUA	1%
	UUG	1%
	CUU	2%
	CUC	3%
	CUA	1%
	CUG	**92%**
Valine	GUU	**60%**
	GUC	2%
	GUA	28%
	GUG	10%
Isoleucine	AUU	16%
	AUC	**84%**
	AUA	0%

http://nar.oxfordjournals.org/content/16/17/8207.short

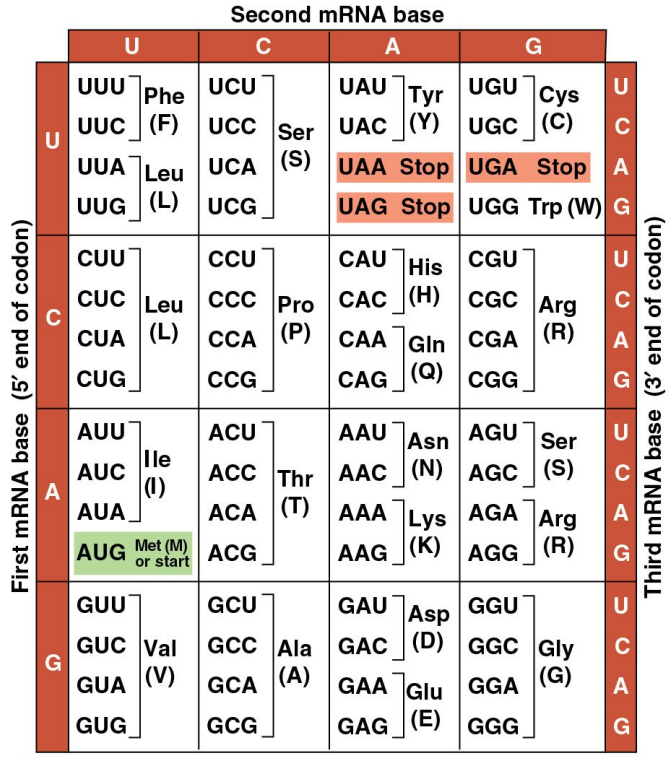


Fig. 17.6, Cambell Biology, 2021 Global edition

MOGELIJKE VERKLARING VOOR CODON BIAS

- Een balans tussen mutatie bias en natuurlijke selectie voor translatie optimalisatie.
  - Snel groeiende organismen zoals *E. coli* en *S. cerevisae* gebruiken optimalisatie van codons die de compositie van hun tRNA reflecteren.
    - Een optimaal codon gebruik zal snellere translatie geven, wat we zien in experimenten.
  - Andere organismen zoals de mens die niet snel groeien of kleine genomen hebben, is er geen codon optimalisatie, maar wordt codon voorkeur door mutatie bias veroorzaakt.

CODON TABELLEN

- De genetische code, of de codon tabel, kan ook variëren.

The following genetic codes are described here:

- 1. The Standard Code
- 2. The Vertebrate Mitochondrial Code
- 3. The Yeast Mitochondrial Code
- 4. The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code
- 5. The Invertebrate Mitochondrial Code
- 6. The Ciliate, Dasycladacean and Hexamita Nuclear Code
- 9. The Echinoderm and Flatworm Mitochondrial Code
- 10. The Euplotid Nuclear Code
- 11. The Bacterial, Archaeal and Plant Plastid Code
- 12. The Alternative Yeast Nuclear Code
- 13. The Ascidian Mitochondrial Code
- 14. The Alternative Flatworm Mitochondrial Code
- 16. Chlorophycean Mitochondrial Code
- 21. Trematode Mitochondrial Code
- 22. Scenedesmus obliquus Mitochondrial Code
- 23. Thraustochytrium Mitochondrial Code
- 24. Rhabdopleuridae Mitochondrial Code
- 25. Candidate Division SR1 and Gracilibacteria Code
- 26. Pachysolen tannophilus Nuclear Code
- 27. Karyorelict Nuclear Code
- 28. Condylostoma Nuclear Code
- 29. Mesodinium Nuclear Code
- 30. Peritrich Nuclear Code
- 31. Blastocrithidia Nuclear Code
- 33. Cephalodiscidae Mitochondrial UAA-Tyr Code

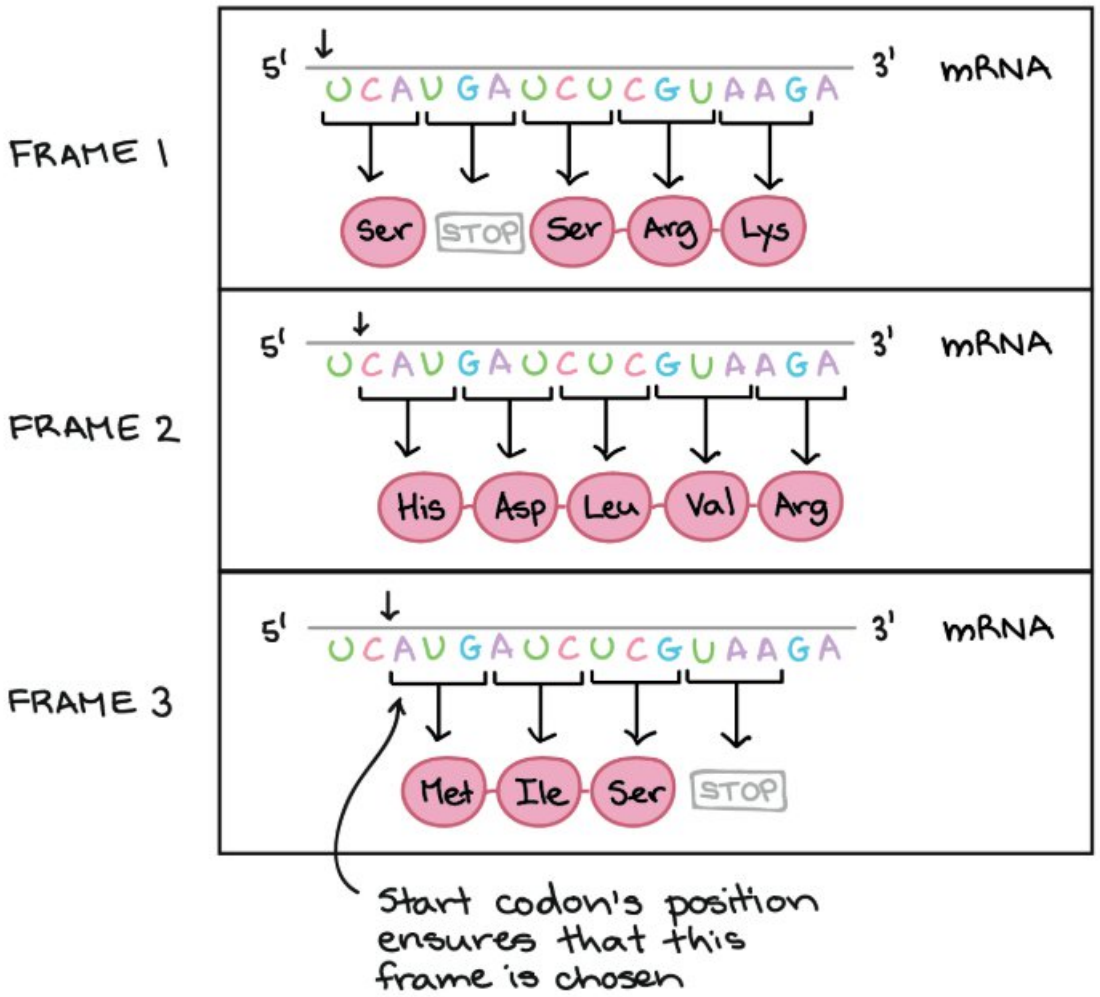
https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

(OPEN) READING FRAMES

- De DNA sequentie tussen start en stop codons.

1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

Open reading frame - Wikipedia



https://cdn.kastatic.org/ka-perseus-images/ed3bfd85b8ec88f74515e63649b9dcd5c976e21e.png

HOE VIND JE GENEN?

- Open readings frames (ORFs)
- Bepaalde lengte van ORFs
  - ORF minimale lengte, bijv. 100 of 150 bp
- Codon gebruik en bias

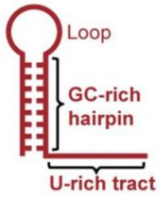
HOE VIND JE GENEN?

Eukaryoten

- Individuele promoteren reguleren genen.
- Intronen in de ORF start-stop definitie van een ORF alleen op spliced mRNA.
- Kozak sequentie:
  - Consensus sequentie
  - Ribosoom assembly en translatie initiatie.
  - Essentieel voor correcte translatie.
  - 5’-(gcc)gccRccAUGG-3’
- Ribosomaal binding locatie (RBS) op mRNA: 5’cap.
- CpG eilanden

Prokaryoten

- Operon gereguleerde gen clusters.
  - Vaak functioneel gerelateerde genen onder controle van een enkele promotor.
- Geen intronen in ORF
- Shine-Delgarno sequentie:
  - Ribosomaal binding locatie (RBS) op mRNA van ~8 base upstream van de start codon voor eiwit synthese/translatie.
  - AGGAGG
- Stop codon gevolgd door terminatie signaal: loop + UUU...



HAN_UNIVERSITY
OF APPLIED SCIENCES

HOE VIND JE GENEN?

Eukaryoten

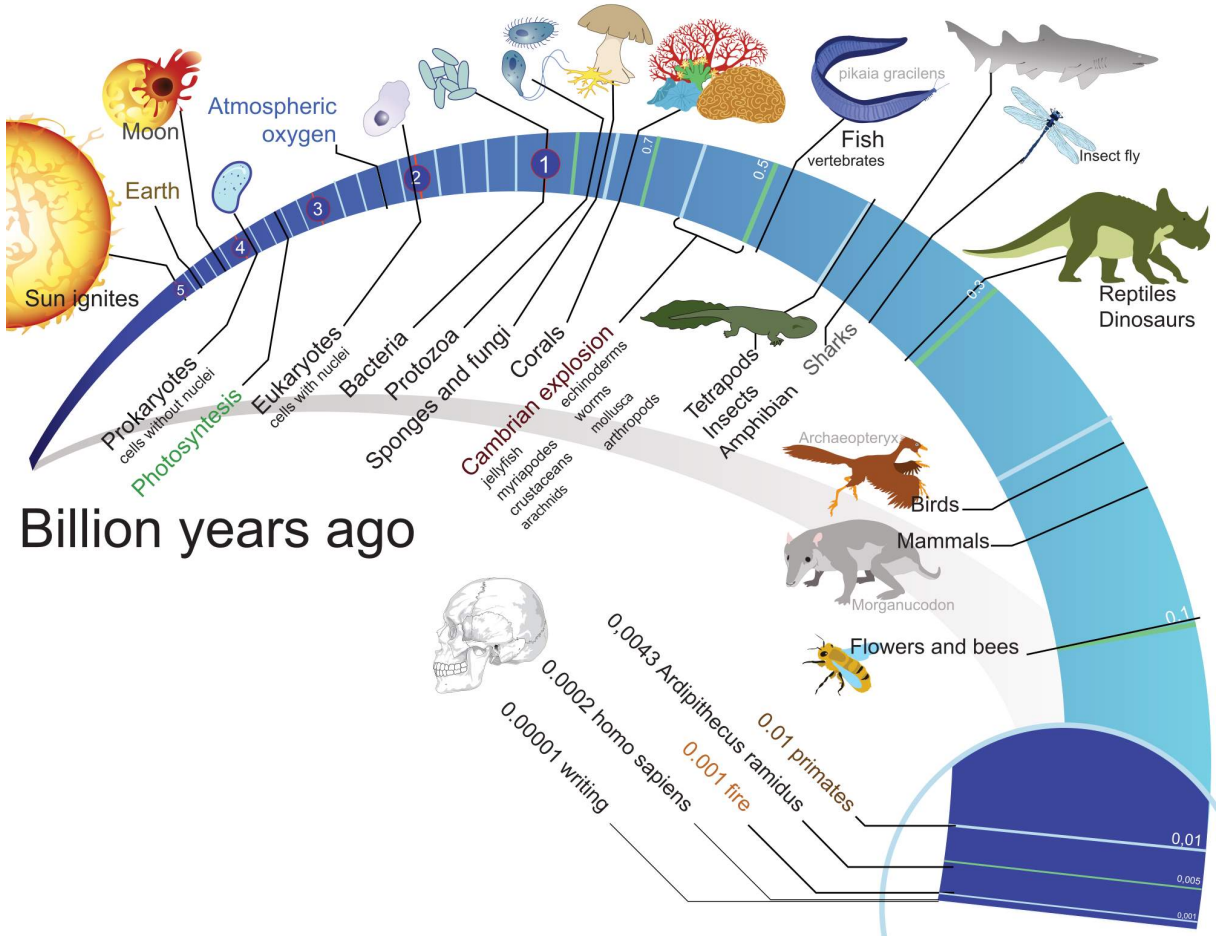
- Poly-A staart
- Splice sites: GT-AG
  - 5’site intron GTAAGT
  - 3’site intron $(Py)_{12}NCAG$

Prokaryoten

NEXT: EVOLUTIE

“The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music.”

- Lewis Thomas



Is Evolution "just a theory"? - Our Planet (ourplnt.com)

EVOLUTIE

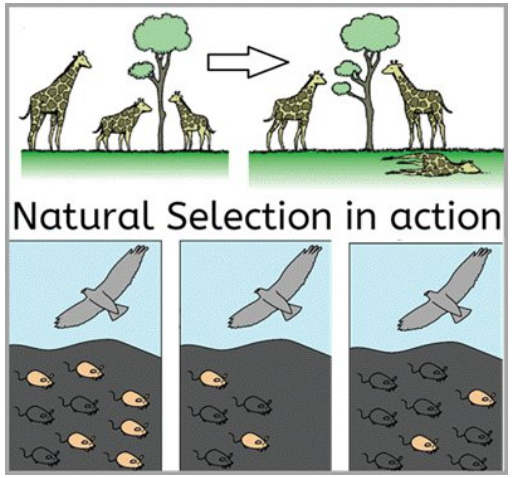
“Nothing in Biology makes sense except in the light of evolution”
- *Theodosius Dobzhansky* (1900-1975)

“Nothing in **Bioinformatics** makes sense except in the light of Biology (and hence evolution)”

(Almost) everything we do in bio-informatics is tied to evolution…

EVOLUTIE, WAT IS HET?

- Evolutie is (geleidelijke) verandering van populaties door overerving van kernmerken en eigenschappen door middel van het genetische materiaal.
- Door middel van genen erft een organisme eigenschappen en kenmerken.
- Mutaties kunnen als gevolg hebben dat er nieuwe eigenschappen ontstaan.
  - Is een mutatie voordelig, biedt het een organisme (betere) kans tot overleven. Dit noemen we natuurlijke selectie.
  - Is een mutatie nadelig, dan zal door natuurlijke selectie een mutatie snel verdwijnen uit een populatie.
  - Voor- of nadelig zijn hangt af van de levensomstandigheden.
- Met genoeg mutaties ontstaat een nieuw soort.



https://www.solpass.org/science6-8-new/s7/images-life/Natural-Selection-Giraffes.png

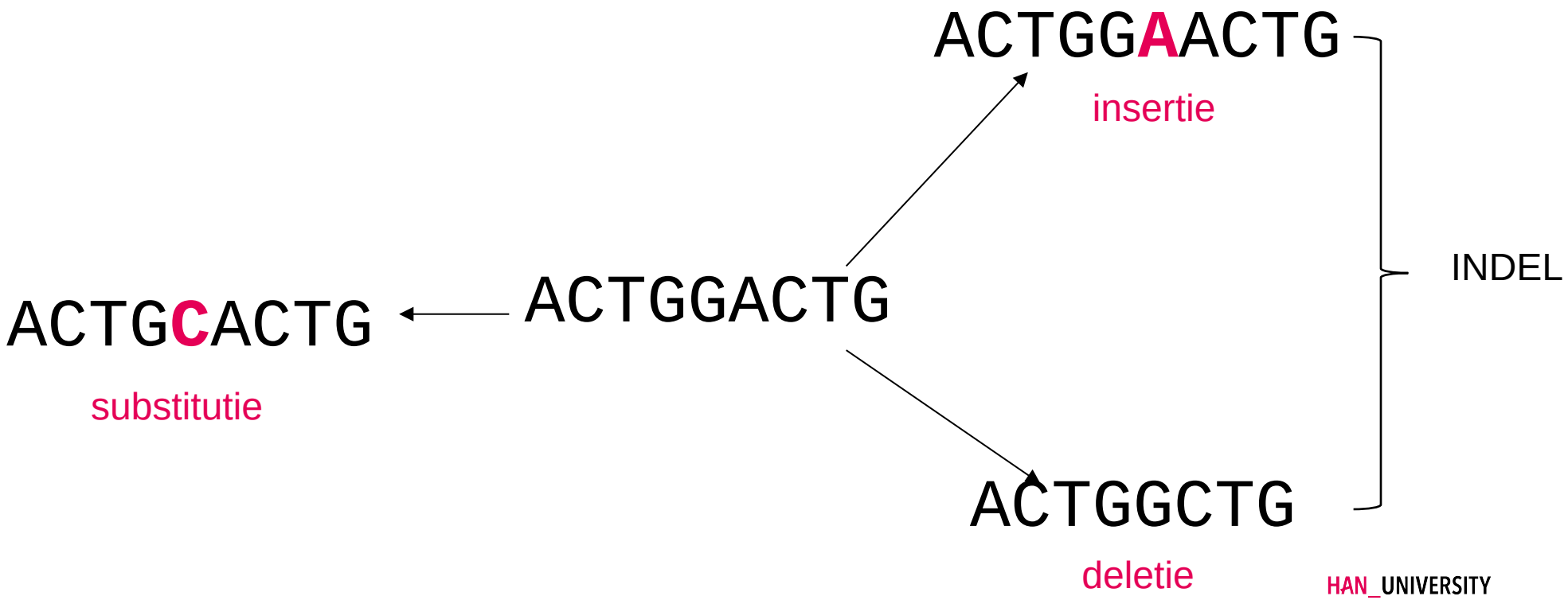
Evolutie (biologie) - Wikipedia

VERANDERINGEN IN DNA

- Mutaties zijn willekeurig
  - Somatisch: niet in gameten en dus niet overerfbaar.
  - Gametisch: in de gameten dus wel overerfbaar.

- Oorzaken van mutaties:
  - Spontaan
    - Replicatie fouten in het DNA (die vervolgens niet hersteld kunnen worden)
  - Omgeving:
    - Geïntroduceerd door externe factoren zoals straling, chemicaliën, virussen of experimentele handelingen.
- Recent onderzoek (2022) beweerd dat mutaties niet spontaan zijn maar onder de invloed van epigenetica! Mutation bias reflects natural selection in Arabidopsis thaliana | Nature

MUTATIES VAN ENKELE BASE (PUNTMUTATIES)



HAN_UNIVERSITY
OF APPLIED SCIENCES

GEVOLGEN VAN PUNTMUTATIES

- Stille mutatie: nieuwe codon codeert voor dezelfde aminozuur. Geen gevolg.
- Missense mutatie: nieuwe codon codeert voor andere aminozuur.
- Nonsense mutatie: nieuwe codon codeert voor een stop codon. Eiwit breekt.
- Frameshift van reading frame bij INDELs: niet werkend eiwit of heel raar eiwit.

ACTGGACTG
ACTGGAACTG insertie
ACTGGCTG deletie

Missense mutation
Sickelcelanemie

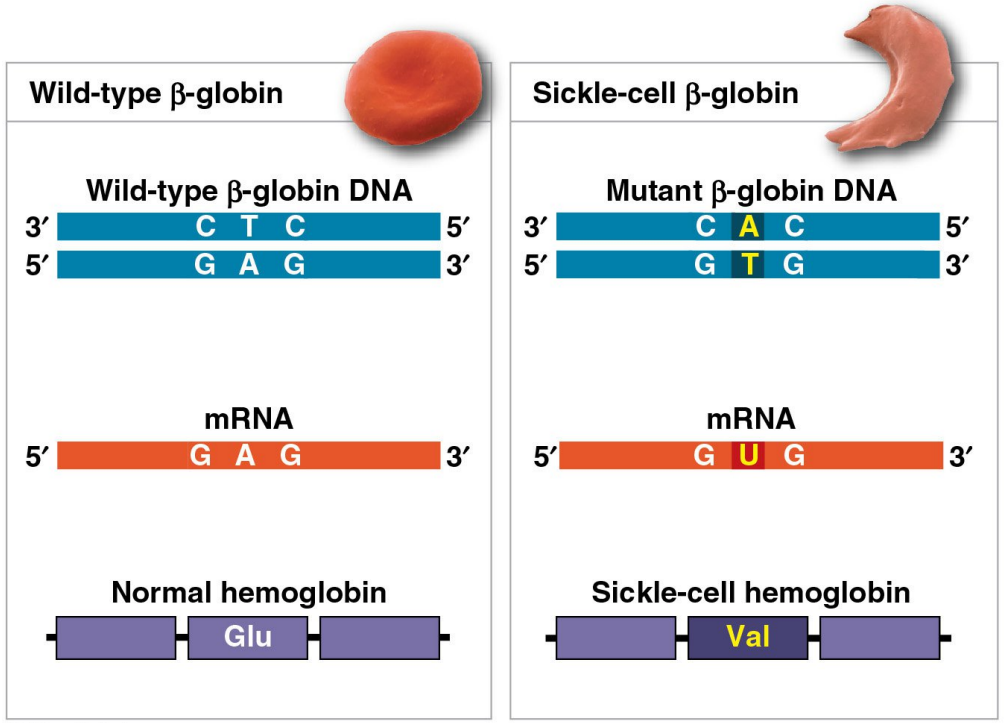
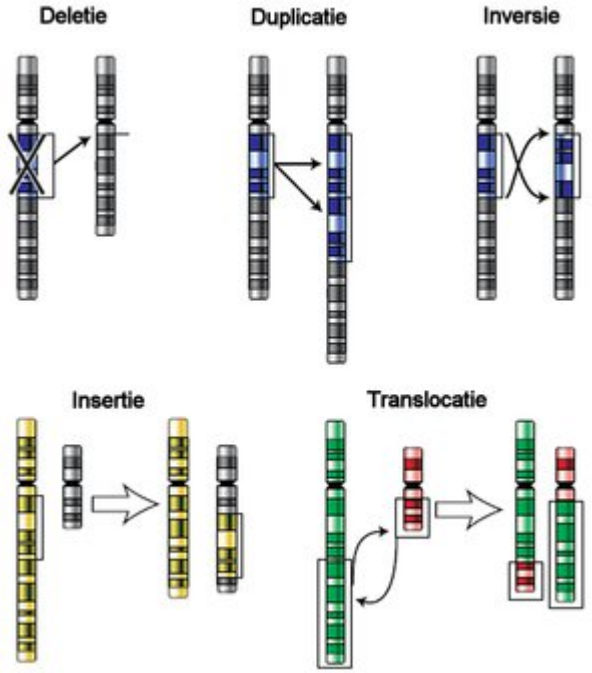


Fig. 17.26, Cambell Biology, 2021 Global edition

MUTATIES VAN MEER DAN ENKELE BASEN

- Segment (chromosomale) mutaties
  - Grote stukken op een chromosoom
- Ploïdie mutaties
  - Aneuploïdie: te weinig of teveel chromosomen.
  - Euploïdie: vermeerdering of vermindering van aantal chromosomen.

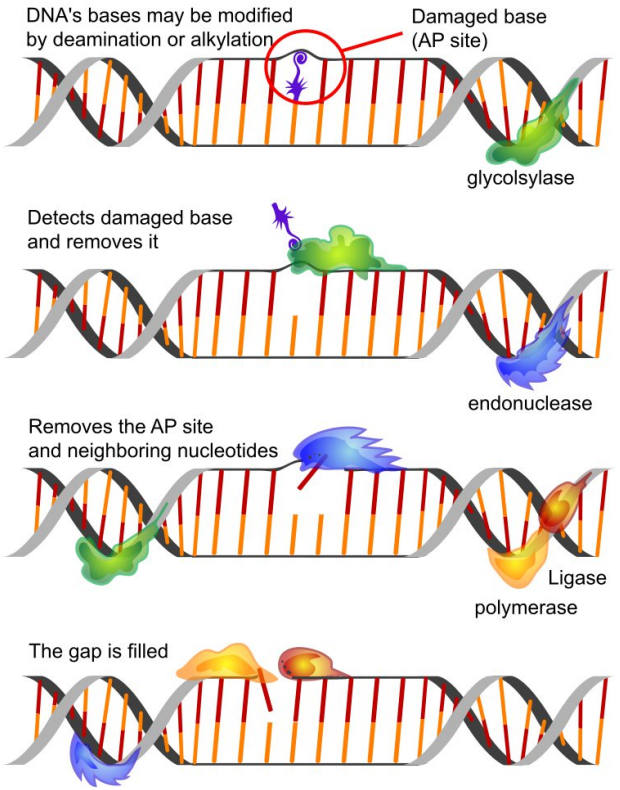


Mutatie (biologie) - Wikipedia

REPARATIE VAN MUTATIES

- DNA replicatie kan foutjes veroorzaken
  - De meeste van deze fouten worden ge-proofread en gecorrigeerd door de enzym DNA polymerase tijdens elongatie.
- Mutaties kunnen ook optreden door externe factoren.
  - AP = abasic site no purine or pyrimidine base door beschadiging.

Base excision repair pathway



Mutation and DNA repair | Brilliant Math & Science Wiki

MEER INFO OVER MUTATIES

- Chromosomal mutation - Definition and Examples - Biology Online Dictionary

EVOLUTIE VAN GENETISCHE CODE

- De genetische code is universeel
  - De codon CCG bijv. codeert voor proline in elk organisme die tot nu toe onderzocht is.
  - Hierdoor kan een gen uit een organisme getransplanteerd, getranscribeerd en dan getransleerd worden in een andere organisme.

(a) Tobacco plant expressing a firefly gene

(b) Mosquito larva expressing a jellyfish gene

Fig. 17.7, Cambell Biology, 2021 Global edition

OORSPRONG VAN GENETISCHE CODE

- Drie theorieën over de oorsprong van de genetische code:
  - Random freeze: de code is van willekeurige oorsprong. De eerste tRNA achtige moleculen met een bepaalde willekeurige codon hadden meer affiniteiten met bepaalde aminozuren. Toen er genoeg eiwitten hun eigen code hadden, was het systeem “bevroren”. Een willekeurige verandering zou het systeem breken.
  - Stereochemische affiniteit: de code is een resultaat van de hoge affiniteit van aminozuren met een codon of anti-codon. Dit impliceert dat eerste tRNA achtige moleculen echt moesten matchen met aminozuren.
  - Optimalisering: de code is zo geoptimaliseerd om maximale fitness te krijgen en zo min mogelijk fouten/mutaties op te lopen.
- Veel variaties en combinaties van deze theorieën mogelijk.

Genetic code - Wikipedia

VIRUSSEN EN CODONS

- Virussen en hun genomen hebben gebruiken een variëteit aan synonieme codons.
- Dit ligt aan de virus en gastheer relatie.
- Waarom is het belangrijk dat een virus hetzelfde codon bias heeft als de gastheer?
- Voor een virus is het van belang om snel de eiwitten tot expressie te brengen.
  - Dus codons die vaak gebruikt worden, betekend meer aanwezige tRNA en dus sneller translatie.

VIRUSSEN EN MUTATIES

- Virussen muteren snel
  - Er is geen proofreading.
  - Er is geen glycolsylase.
  - Heel veel nakomelingen.
  - Adaptie tegen immuunsystemen.

EVOLUTIONAIR EN FUNCTIONEEL BELANG VAN INTRONEN

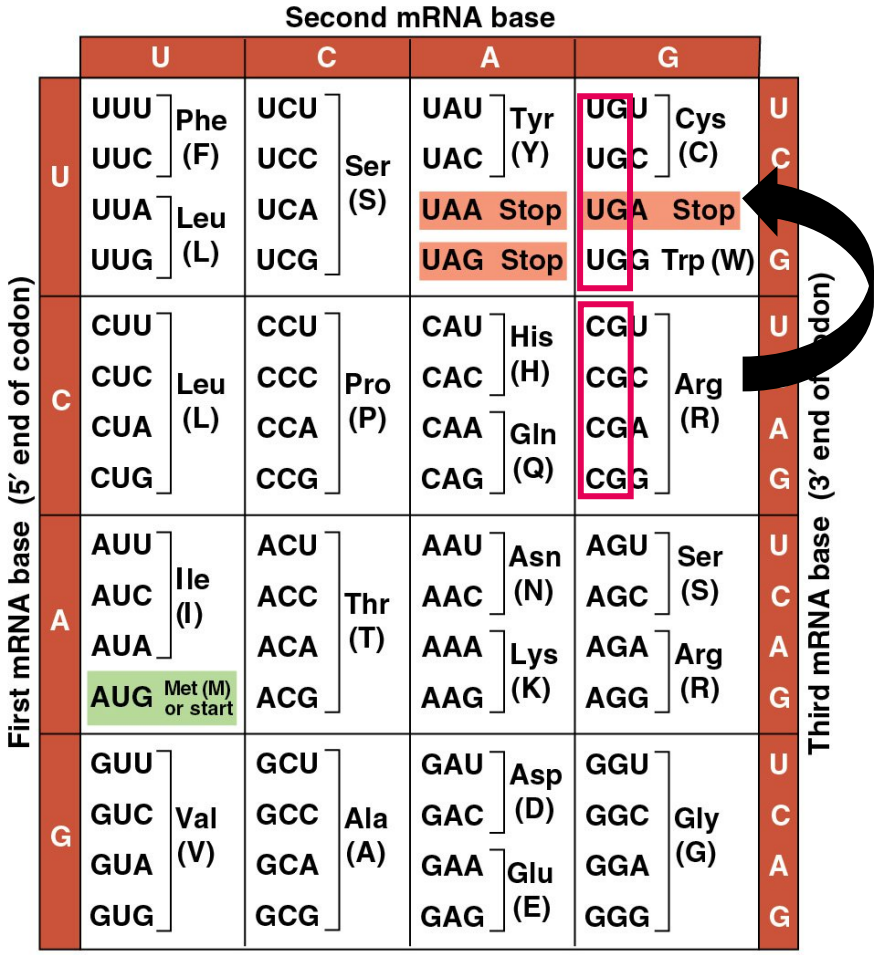
- Geen specifieke functies gevonden in intronen, sommige willen wel een eens product hebben die iets met regulatie van gen expressie te maken hebben.
- Waarschijnlijk hebben intronen veel voordelen in adaptie van organismen.
- Voornamelijk veel invloed op gen producten.
  - Enkele gen kan voor meer producten coderen (alternative splicing).
- Aanwezigheid van intronen in een gen heeft mogelijk gefaciliteerd dat er meer voordelige eiwitten geproduceerd konden worden.

SAMENVATTING

- Codon gebruik = gebruik van de codons in een gen of organisme.
- Codon bias = de voorkeur van een organisme voor het gebruik van bepaalde codons ten opzichte van andere codons.
  - Wordt veroorzaakt door transcriptie optimalisatie
    - Hoe meer tRNA met een bepaalde codon, hoe sneller dat codon gebruikt wordt (kip en ei verhaal), hoe sneller transcriptie kan plaatsvinden.
  - Wordt veroorzaakt door mutatie bias
    - Codons met nucleotiden die sneller muteren liever niet

SAMENVATTING

- CpG hebben met regulatie te maken.
  - C's naast G's kunnen makkelijk gemetyleerd worden.
  - Methylatie zorgt ervoor dat genen minder actief afgeschreven worden.
- Een probleem bij dit (soort) gen regulatie is dat C's die naast een G gemethyleerd worden, makkelijk muteren naar T's.
  - Dit betekend dan we liever minder vaak codons gebruiken die CG combinaties hebben, omdat we dan sneller mutaties in ons eiwitten krijgen.



SAMENVATTING

- We kunnen genen vinden door naar indicaties van genen te kijken (bijv. CpGs, codon bias, GC%, ORFs), en naar bepaalde sequenties te kijken (bijv. promotor sequenties, start/stop codons, splice sites, Kozak en Shine-Delgarno).

SAMENVATTING

- Evolutie wordt teweeg gebracht door veranderingen in genetisch materiaal (DNA) in een populatie.
- Veranderingen in het DNA komen door mutaties.
- Er zijn verschillende soorten mutaties, voor-, nadelig, of zonder verandering.