Academie toegepaste biowetenschappen en chemie

Bio-Informatica Course 3

5. Proteins and alignments

HAN_UNIVERSITY OF APPLIED SCIENCES

Onderzoeksvragen genoomonderzoek

HIV kan bij de mens AIDS veroorzaken.
AIDS=Acquired Immune Deficiency Syndrome
Vooral HIV-1 erg gevaarlijk (pandemie)

I. Waarom is HIV-1 hoog virulent?
II. Wat is de fylogenetische oorsprong van HIV-1?

Overzicht onderzoeksplan 3b

Week	Onderwerp	Activiteiten	
1	Oriëntatie op onderwerp Sequenties verzamelen	Inlezen Stap 1	NCBI database
2	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		

Vandaag

Herhaling vorige week
Uitleg deze week

Deze week

Alignments

Wat is een alignment?

Waarom doen we alignments?

De kwaliteit van alignments.

Alignments en evolutie?

Hfst 3 Pevsner blz 69-78

Sequentie alignments

Wat is het?

Sequentie alignments

Wat is het?

Vergelijken van twee of meer DNA, RNA of eiwit sequenties.

Sequentie alignments

Wat is het?

Vergelijken van twee of meer DNA, RNA of eiwit sequenties.

Waarom?

Sequentie alignments

Wat is het?

Vergelijken van twee of meer DNA, RNA of eiwit sequenties.

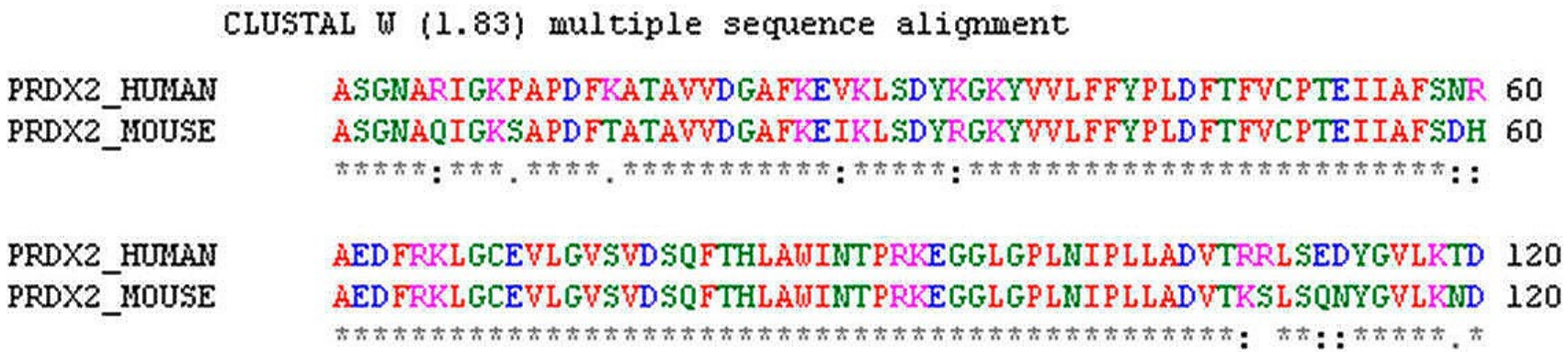
Waarom?

Identificeren van regio's die vergelijkbaar zijn als gevolg van een functionele, structurele of evolutionaire relatie.

Sequentie alignments

Sequenties zo indelen zodat vergelijkbare gedeeltes van de sequentie met vergelijkbare eigenschappen worden uitgelijnd.

- DNA, RNA of eiwit.

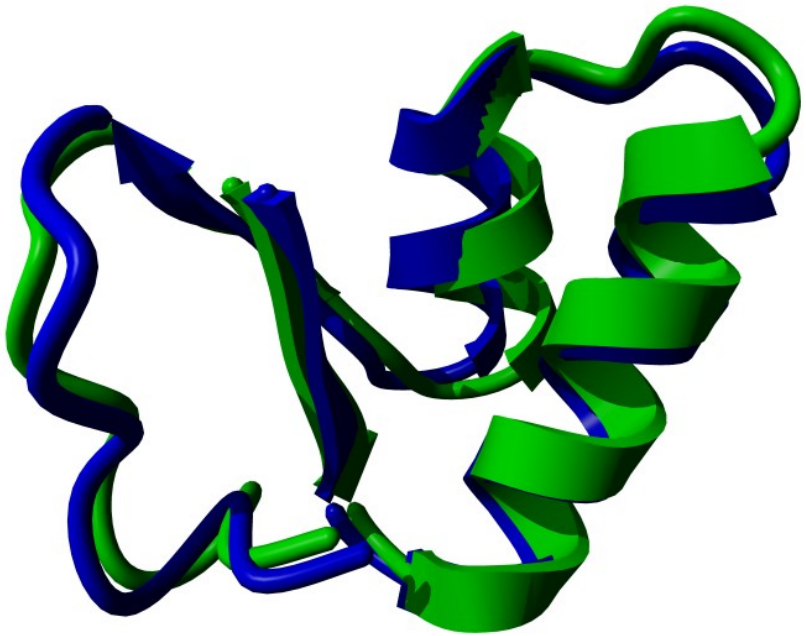


Alignen van eiwitten

Dit zie je:

```
cram_craab: TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
thn_dencl : KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH
```

Maar eigenlijk doe je dit:



Pairwise alignment

Snel en simpel

Niet altijd betrouwbaar

score berekenen aan de hand van identities en similarities van de aminozuren.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                     ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .****************     :  *.: :
```

Pairwise alignment

Identity: Exacte matches
Similarity (positives): vergelijkbare match (bv. R→K)

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                     ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .*****************     :  *.: :
```

Gaps en mismatches

Gaps (-) zijn belangrijk voor het beter alignen.
Ze simuleren indel events.
Een mismatch simuleert een substitutie

```
ATCGAT    ATCGAT    ATCGAT
|           ||||    | ||||
ACGAT      ACGAT    A-CGAT
```

Gaps

Graag hebben we een alignment die evolutionair logisch is.

Wat is logischer?

```
ACGTCTGATACGCCGTATAGTCTATCT
ACGTCTGAT------ATAGTCTATCT

ACGTCTGATACGCCGTATAGTCTATCT
AC-T-TGA--CG-CGT-TA-TCTATCT
```

Een alignment score

De som van alle matches van een alignment, met gap penalties daarvan afgehaald.

Match score: +1

Mismatch score: 0

Opening gap penalty: -2

Lengte penalty: -1

Score?

```
ACGTCTGATACGCCGTATAGTCTATCT
||||-||||     |||-|||||||
ACGTTTGAT------ATACTCTATCT
```

Een alignment score

Match score: +1
Mismatch score: 0
Opening gap penalty: -2
Lengte penalty: -1

Score?

```
ACGTCTGATACGCCGTATAGTCTATCT
||||-||||     |||-|||||||
ACGTTTGAT-------ATACTCTATCT
```

Score = 18-(2+6)=10

Scoring systemen – PAM250

Bij het vergelijken van eiwitsequenties

- sommige aminozuren lijken meer op elkaar dan andere.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Scoring systemen – PAM250

Bij het vergelijken van eiwitsequenties

- sommige aminozuren lijken meer op elkaar dan andere.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

```
CSTPAG
|+  |+|
CTWPGG
```

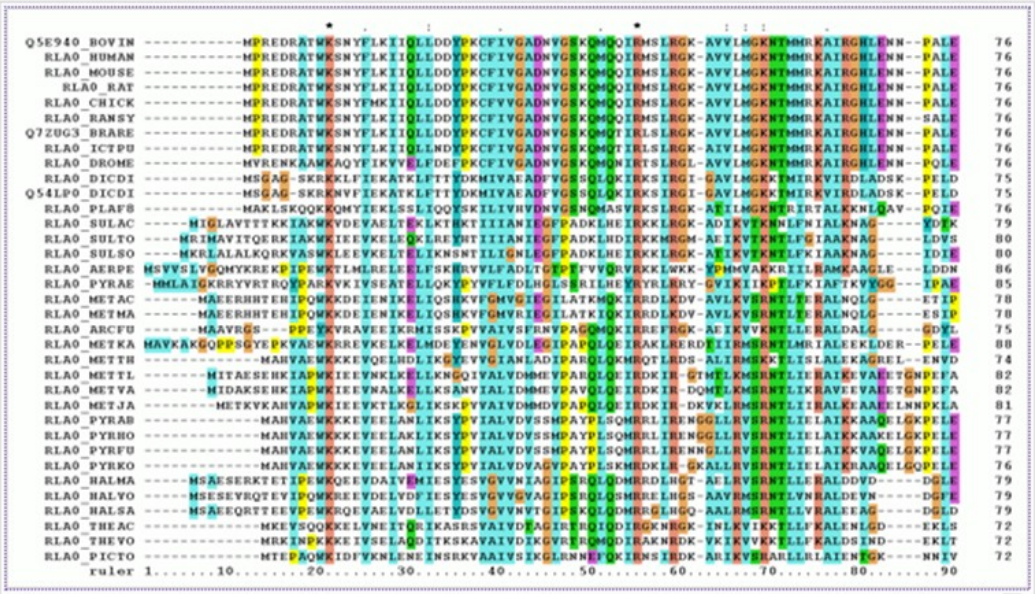
Score=
12+1-5+6+1+5=20

Multiple sequence alignment

Zwaar voor de computer

- Iedere sequentie met iedere andere sequentie alignen.
- Scores kunnen gebruikt worden voor een fylogenetische boom.

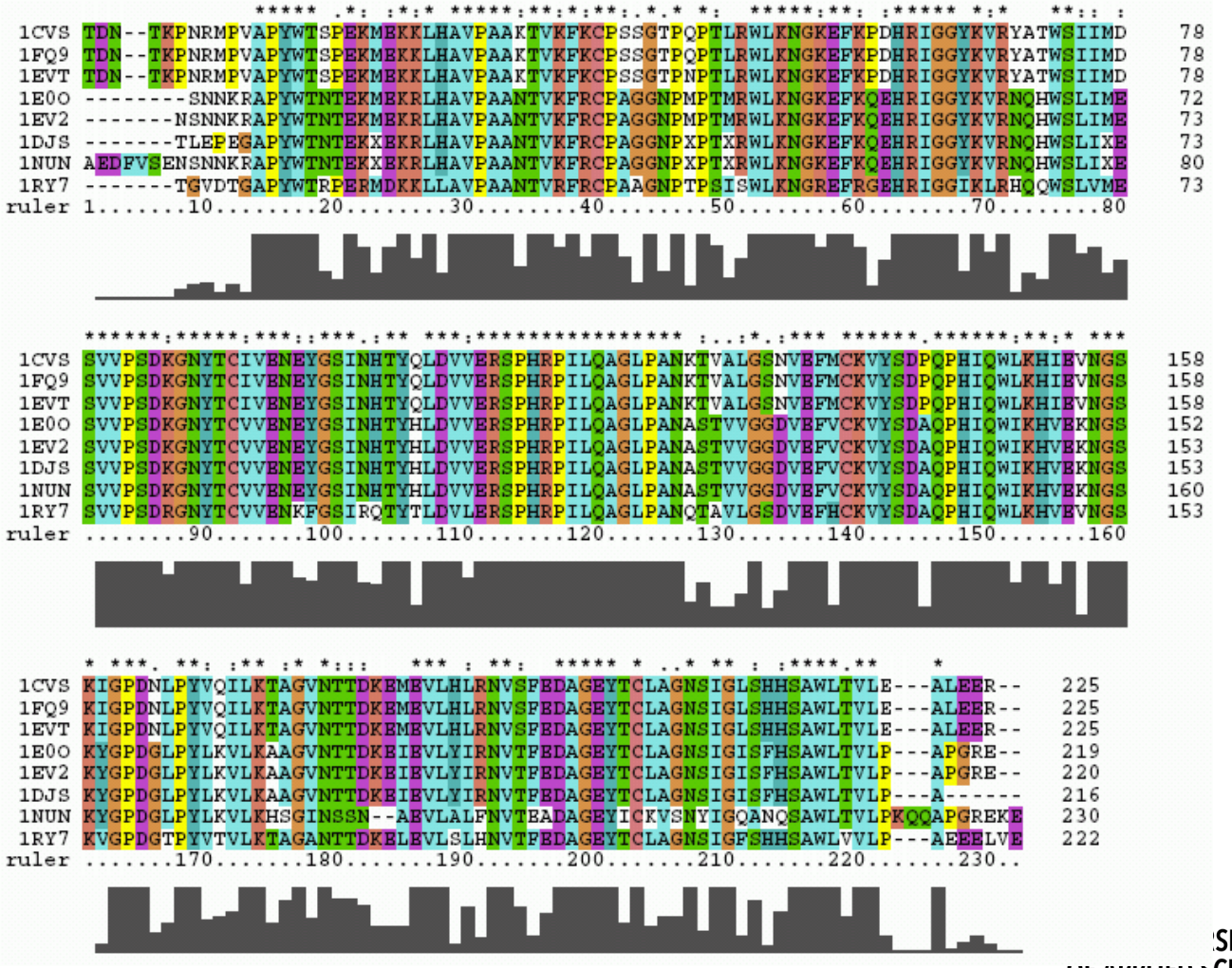
Betrouwbaarder: Meer informatie beschikbaar.



First 90 positions of a protein multiple sequence alignment of instances of the acidic ribosomal protein P0 (L10E) from several organisms. Generated with ClustalX.

HAN_UNIVERSITY OF APPLIED SCIENCES

Multiple sequence alignment



Multiple sequence alignment

Meer sequenties meer informatie

CTGAGCGACGTAGCGCTCTTCGAGC

Multiple sequence alignment

Meer sequenties meer informatie

```
CTGAGCGACGTAGCGCTCTTCGAGC
CTGAACGATTTAGCCATTTTCGAGC
```

Multiple sequence alignment

Meer sequenties meer informatie

```
CTGAGCGACGTAGCGCACTTCGAGC
CTGAATGATTTAGGCATTTTCGAAT
ATGCACTACTTAGTGATGGTCGTGT
ATGCATGATGTAGCGATGGTCGACC
```

Waarom alignen?

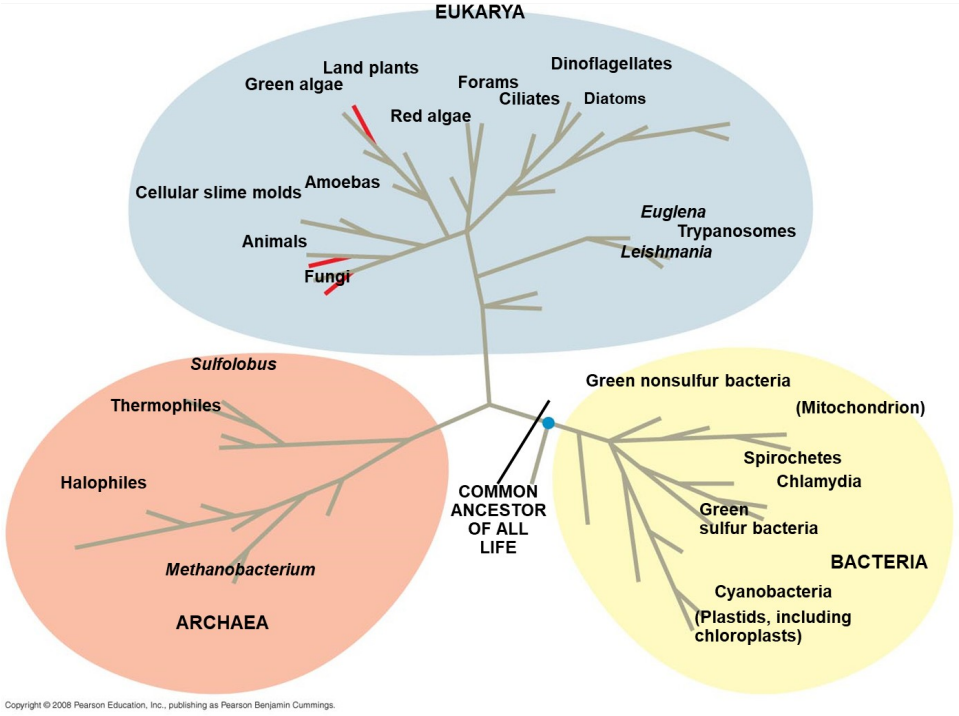
Homologen vinden:

- Evolutionaire events identificeren.
  - Duplicaties
  - Soortvorming

Extrapoleren van kennis over de ene sequentie naar een andere sequentie.

Waarom alignen?

Structuur en functie van een eiwit ontrafelen
Identificatie van geconserveerde sequenties.
Fylogenie
Identificeren van homologen, orthologen en paralogen



Homologen, paralogen en orthologen

Homologe sequenties

- Hebben een gemeenschapelijk voorouder**sequentie**

Homologen, paralogen en orthologen

“Muizen en ratten myoglobine zijn voor 98% homoloog.”

Homologen, paralogen en orthologen

“Muizen en ratten myoglobine zijn voor 98% homoloog.”

ONJUIST

Ze zijn homoloog of ze zijn het niet.

Homologen, paralogen en orthologen

Homologe sequenties

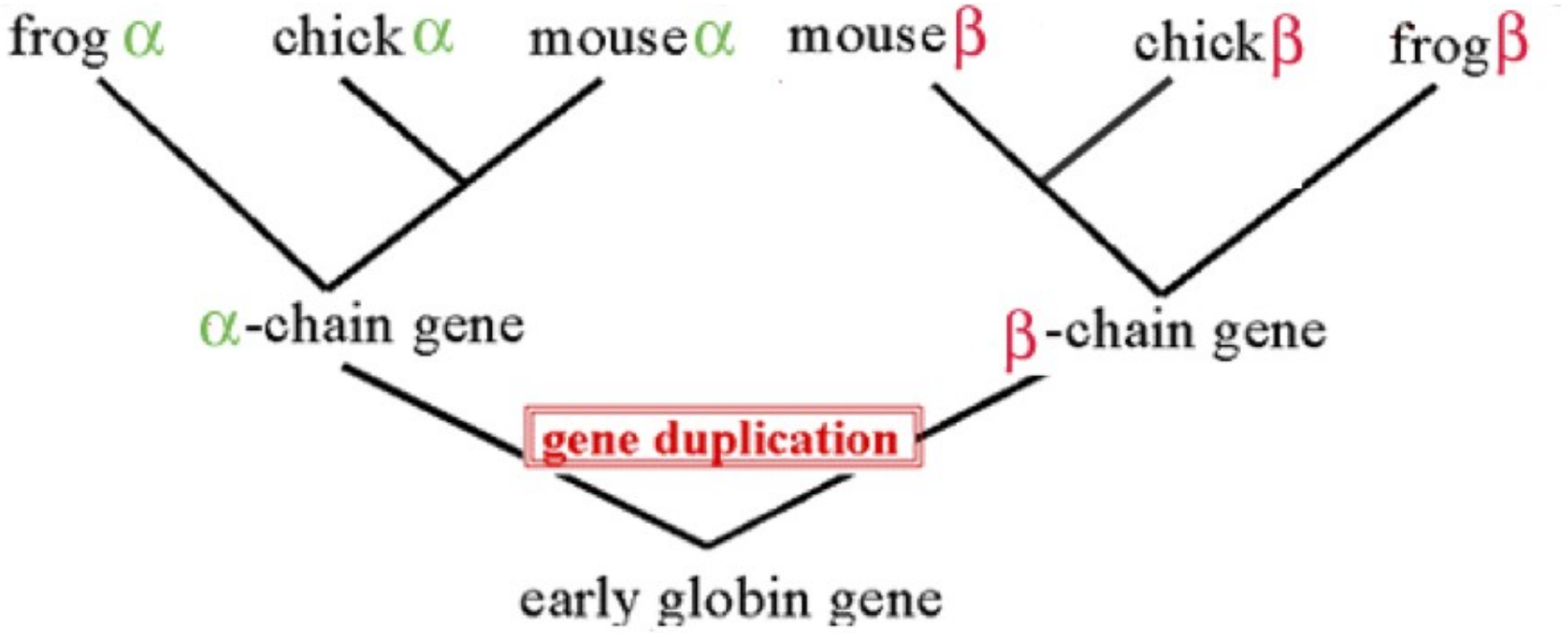
Waarschijnlijk homoloog als:

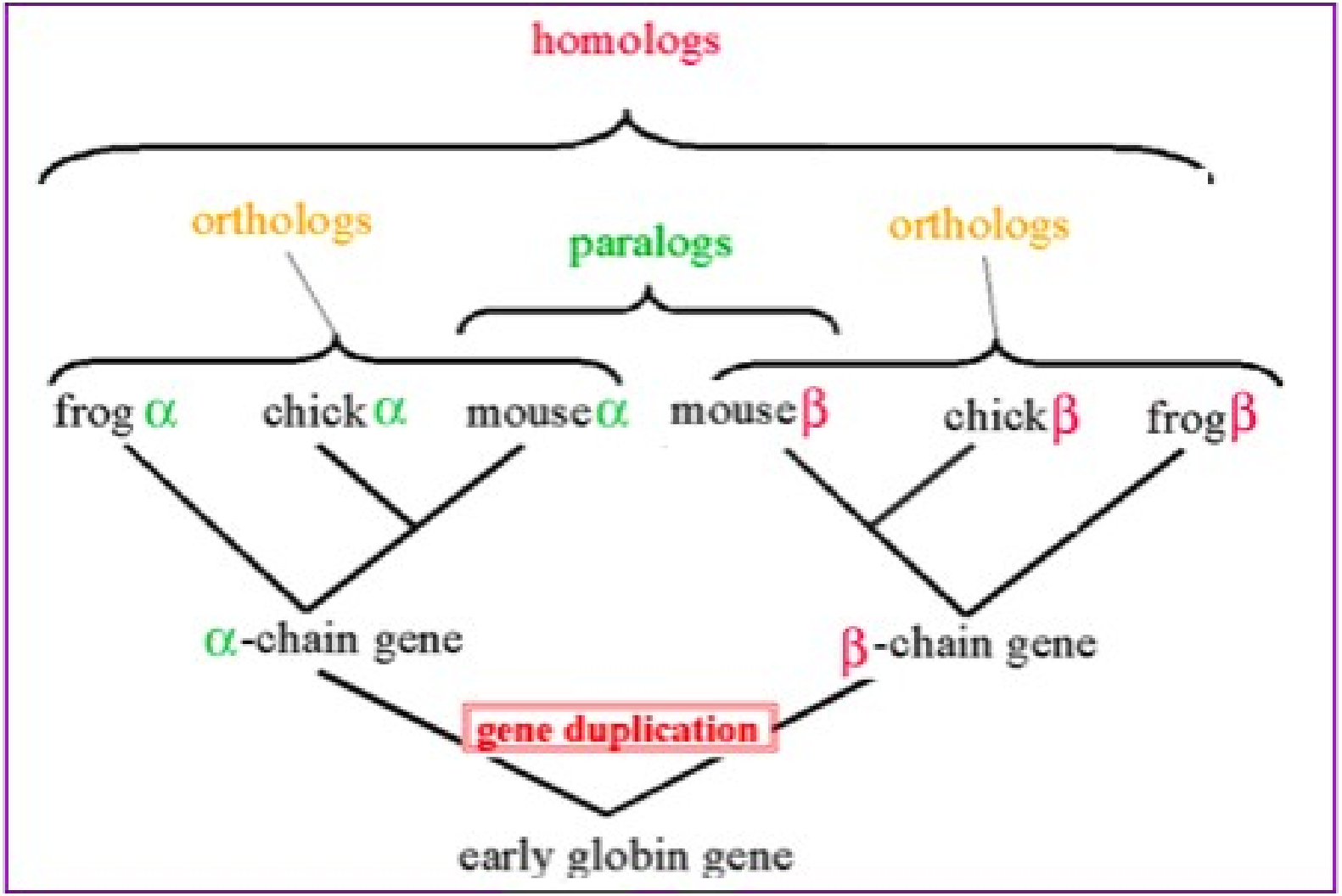
- 2 eiwitsequenties (>100 a.z.) >25% identiteit
- 2 DNA sequenties (>100 nt) >70% indentiteit

Maar…

Bij een lagere identiteit kan er nog steeds sprake zijn van homologie.

Orthologs & Paralogs





http://www.thegreatgoodplace.com/tt/attach/1/1205352441.gif

BLAST

BLAST

- Onderzoeker kan een query sequentie vergelijken met een database van sequenties, en daarmee vergelijkbare sequenties identificeren.

Identificeren van paralogen en orthologen

Samenvattend

Alignen = Vergelijken van twee of meer DNA of eiwit sequenties.

Score matrix gebruiken we om de alignment te kwantificeren.

Alignen gebruiken we om orthologen en paralogen te vinden en om onze informatie over een sequentie te extrapoleren naar een andere sequentie.

Bronnen

Afbeeldingen afkomstig van:

Campbell – Biology A global Approach. 10/11th edition, Uitgever: Pearson
(Verplicht op de boekenlijst van de opleiding)

Lehninger- Principles of biochemistry, fifth edtion. 2008 Uitgever W.H. Freeman & co Ltd

http://www.thegreatgoodplace.com/tt/attach/1/1205352441.gif