

HK07 – Les 2
Sequentiealignering

Yves Moreau
3de jr. Burg. Ir. Elektrotechniek
Dataverwerking & Automatisatie
2001-2002

Practische regelingen

- Cursus
  - Transparanten
  - Gedetailleerde lijst van hoofdstukken en aanvullende artikels
- Oefenzitting
  - Afgeven van een kort verslag op het einde van iedere zitting
- Les 1 : *Essential Cell Biology*
- Les 2 : Hoofdstuk 2 van *Biological sequence analysis*

Overzicht

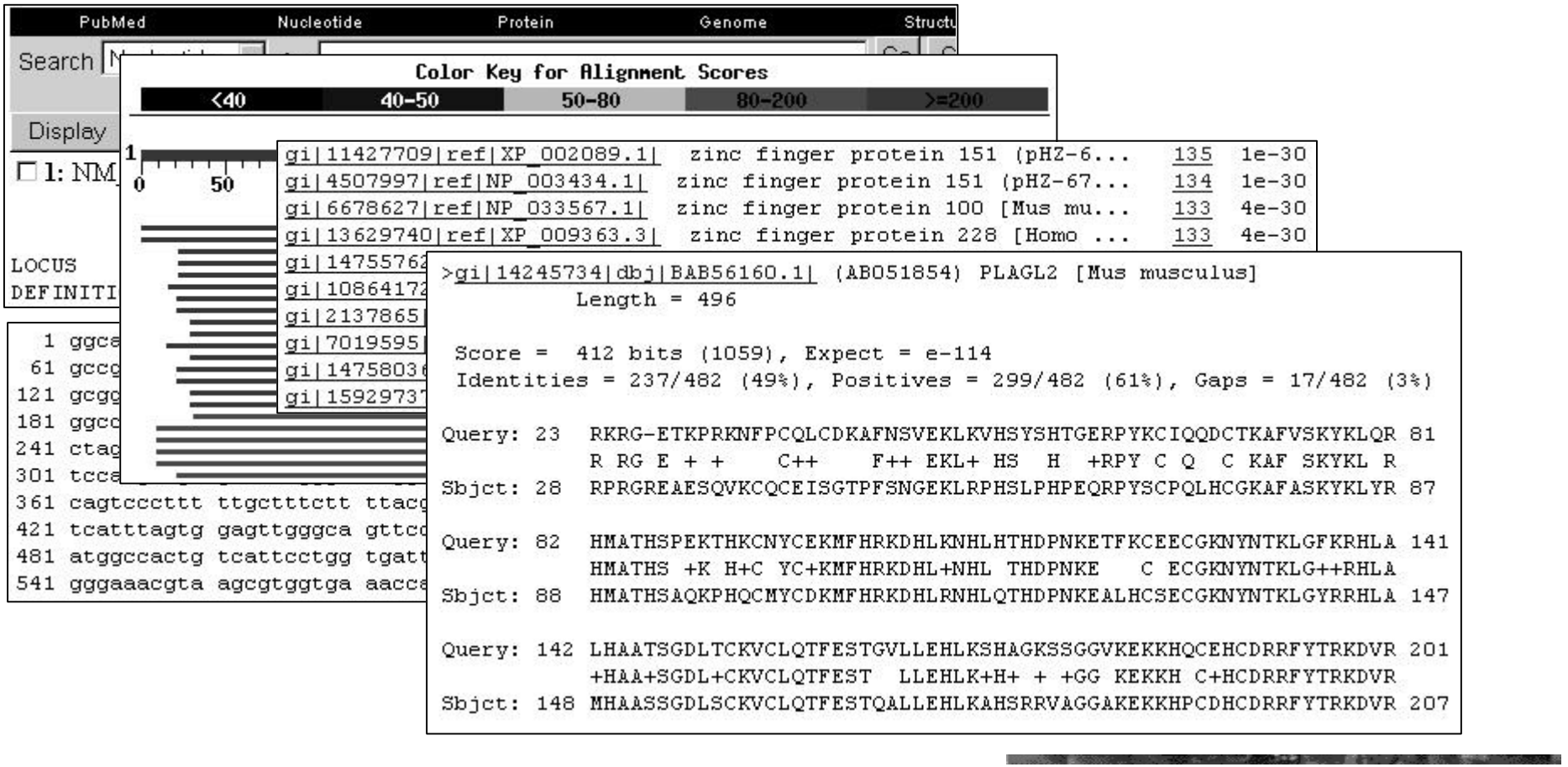
- Aligneren van twee sequenties
  - DNA
  - Proteïnen
- Similariteit vs. homologie
  - Similariteit
  - Homologie
    - Orthologie
    - Paralogie
- Elementen van een alignering
- Dynamisch programmeren

Overzicht

- Globale alignering
  - Needleman-Wunsch algoritme
- Lokale alignering
  - Smith-Waterman algoritme
  - Affiene gapkost
- Substitutiematrices
  - PAM
  - BLOSUM
  - Gapkost
- Significantieberekening
- BLAST

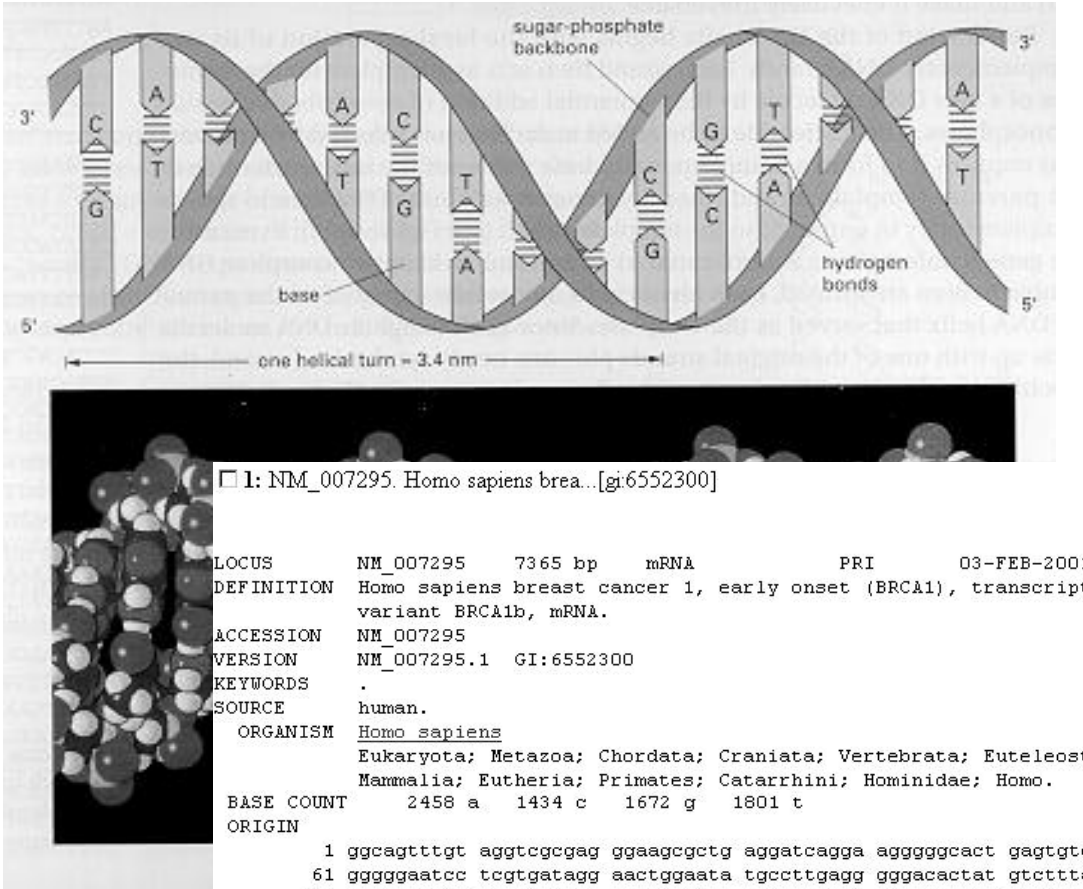
Biologische basis voor alignering

BLAST for discovery



DNA

- DNA helix
- Complementaire strengen (A-T, G-C)
- Georienteerd van 5’ naar 3’



1: NM_007295. Homo sapiens brea...[gi:6552300]

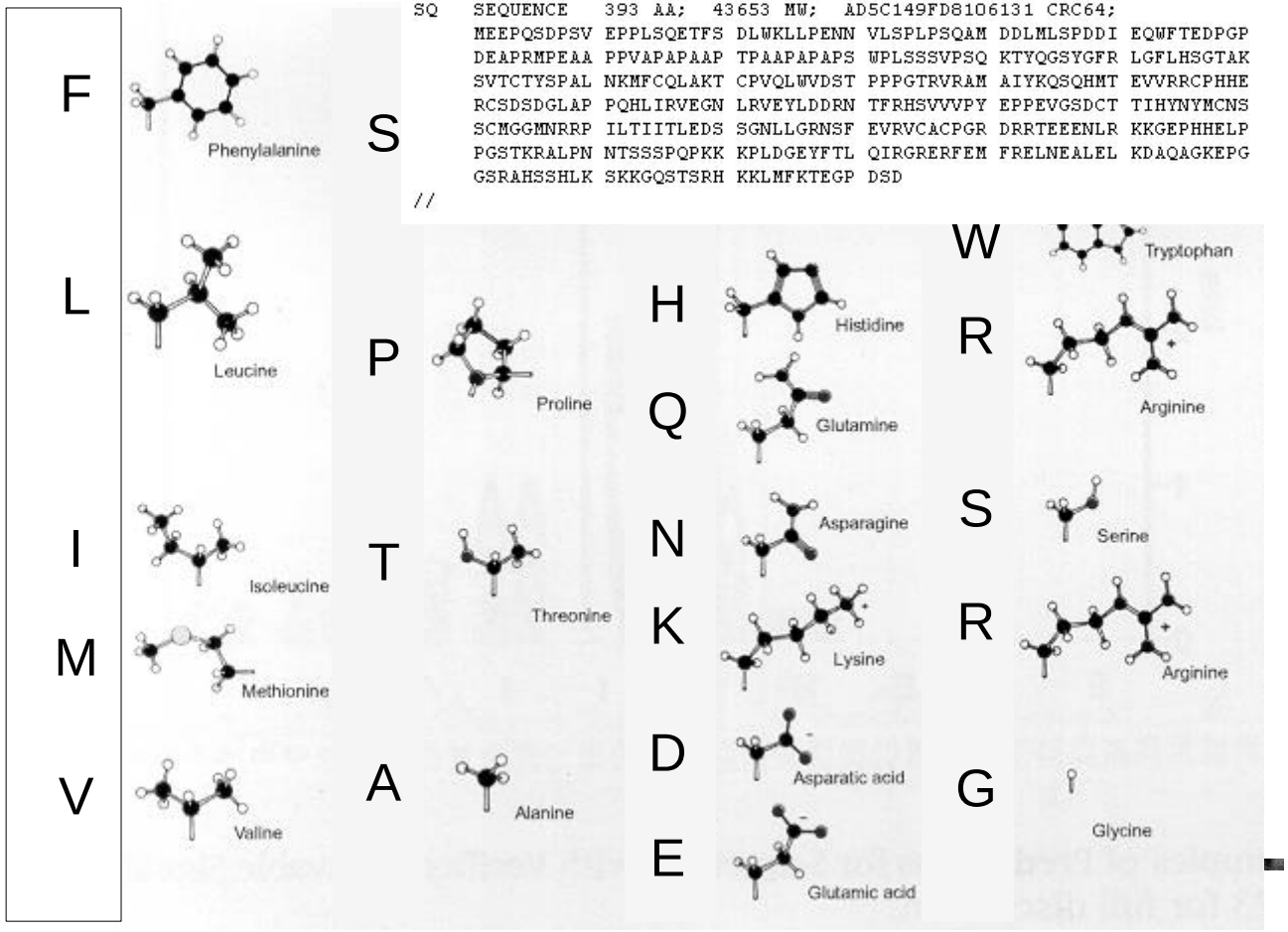
```
LOCUS       NM_007295    7365 bp    mRNA            PRI       03-FEB-2001
DEFINITION  Homo sapiens breast cancer 1, early onset (BRCA1), transcript
            variant BRCA1b, mRNA.
ACCESSION   NM_007295
VERSION     NM_007295.1  GI:6552300
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
 BASE COUNT     2458 a   1434 c   1672 g   1801 t
ORIGIN
        1 ggcagtttgt aggtcgcgag ggaagcgctg aggatcagga agggggcact gagtgtccgt
       61 gggggaatcc tcgtgatagg aactggaata tgccttgagg gggacactat gtctttaaaa
      121 acgtcggctg gtcatgaggt caggagttcc agaccagcct gaccaacgtg gtgaaactcc
      181 gtctctacta aaaatacaaa aattagccgg gcgtggtgcc gctccagcta ctcaggaggc
      241 tgaggcagga gaatcgctag aacccgggag gcggaggttg cagtgagccg agatcgcgcc
      301 attgcactcc agcctgggcg acagagcgag actgtctcaa aacaaaacaa aacaaaacaa
      361 aacaaaaaac accggctgtt cattggaaca gaaagaaatg gatttatctg ctcttcgcgt
      421 tgaagaagta caaaatgtca ttaatgctat gcagaaaatc ttagagtgtc ccatctgtct
```

Proteïne

- Groot polymeer op basis van
  20 aminozuren

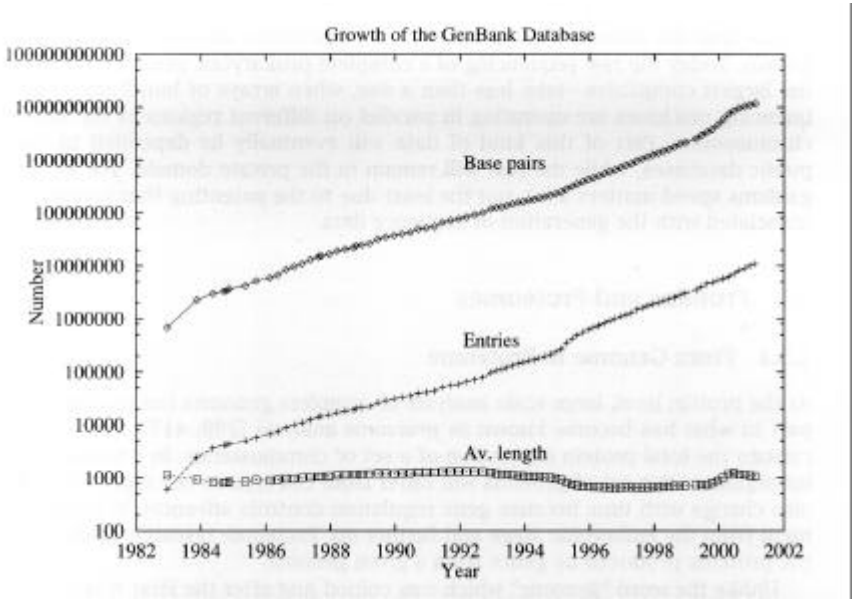
SWISSPROT:P53_HUMAN

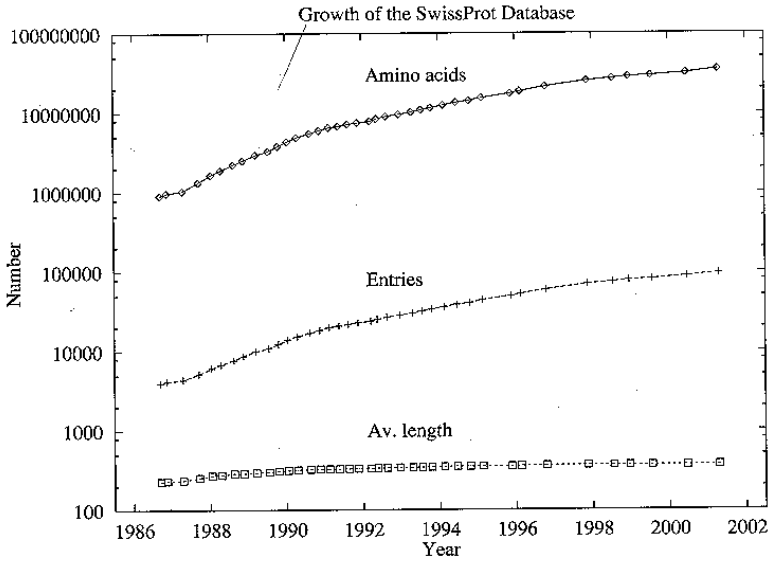
```
ID   P53_HUMAN      STANDARD;      PRT;   393 AA.
AC   P04637; Q16848; Q9UBI2;
DT   13-AUG-1987 (Rel. 05, Created)
DT   01-MAR-1989 (Rel. 10, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53) (PHOSPHOPROTEIN
DE   P53).
GN   TP53 OR P53.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI TaxID=9606;
SQ   SEQUENCE   393 AA;  43653 MW;  AD5C149FD8106131 CRC64;
     MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI EQWFTEDPGP
     DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPLSSSVPSQ KTYQGSYGFR LGFLHSGTAK
     SVTCTYSPAL NKMFCQLAKT CPVQLWVDST PPPGTRVRAM AIYKQSQHMT EVVRRCPHHE
     RCSDSDGLAP PQHLIRVEGN LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS
     SCMGGMNRRP ILTIITLEDS SGNLLGRNSF EVRVCACPGR DRRTEEENLR KKGEPHHELP
     PGSTKRALPN NTSSSPQPKK KPLDGEYFTL QIRGRERFEM FRELNEALEL KDAQAGKEPG
     GSRAHSSHLK SKKGQSTSRH KKLMFKTEGP DSD
//
```



Evolutie van sequentiedatabanken

- Genbank
- SWISSProt





Genomen

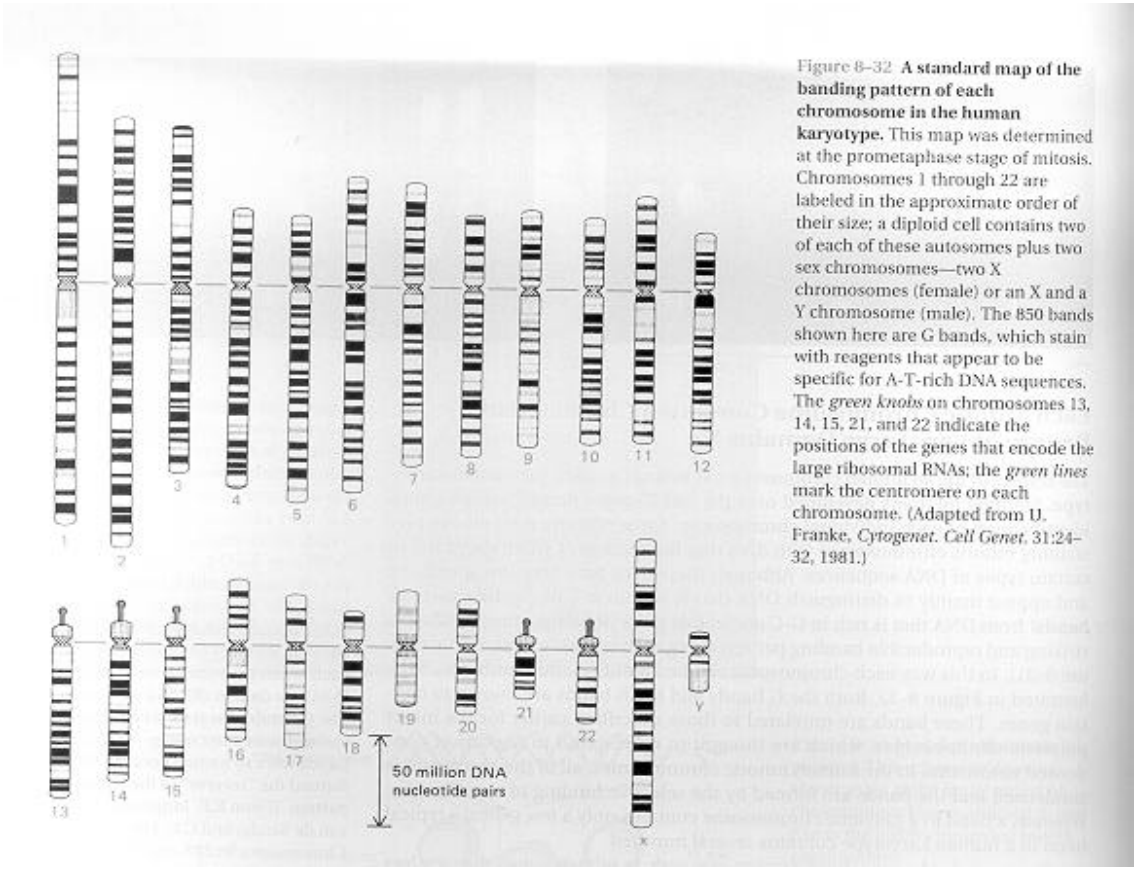
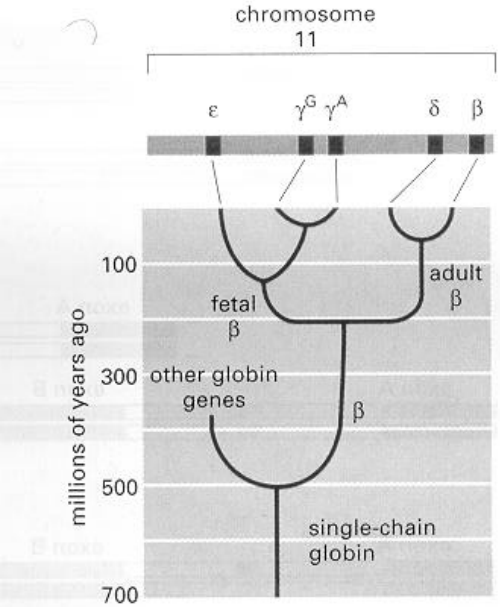


Figure 8–32 **A standard map of the banding pattern of each chromosome in the human karyotype.** This map was determined at the prometaphase stage of mitosis. Chromosomes 1 through 22 are labeled in the approximate order of their size; a diploid cell contains two of each of these autosomes plus two sex chromosomes—two X chromosomes (female) or an X and a Y chromosome (male). The 850 bands shown here are G bands, which stain with reagents that appear to be specific for A-T-rich DNA sequences. The *green knobs* on chromosomes 13, 14, 15, 21, and 22 indicate the positions of the genes that encode the large ribosomal RNAs; the *green lines* mark the centromere on each chromosome. (Adapted from U. Franke, *Cytogenet. Cell Genet.* 31:24–32, 1981.)

Moleculaire evolutie

- Genomen evolueren door imperfecte replicatie en natuurlijke selectie
- Genduplicaties creëren families van genen



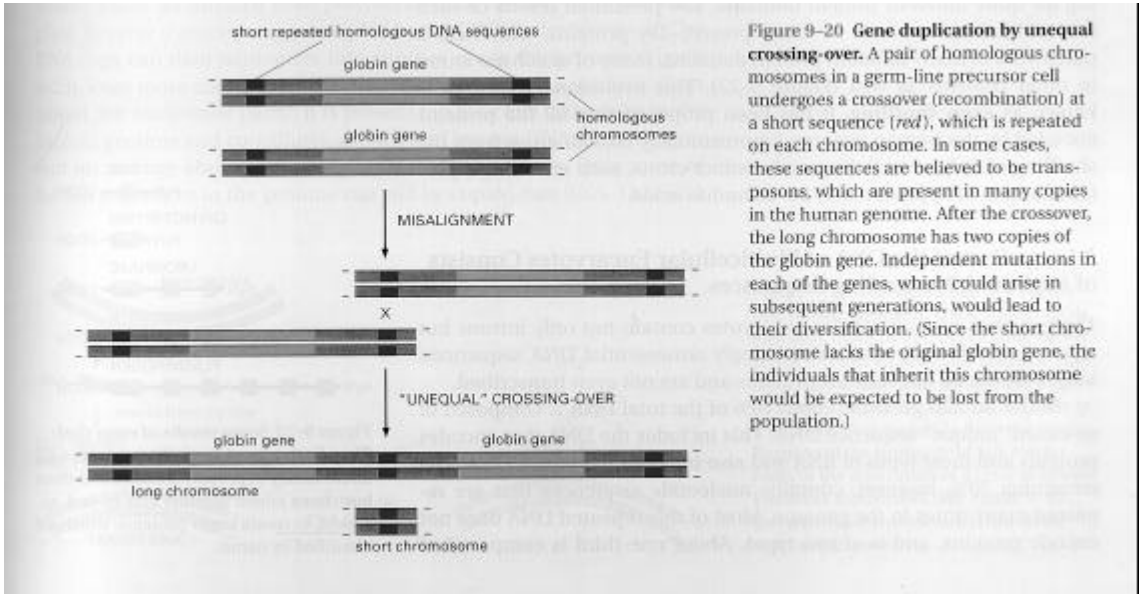


Figure 9–20 **Gene duplication by unequal crossing-over.** A pair of homologous chromosomes in a germ-line precursor cell undergoes a crossover (recombination) at a short sequence (*red*), which is repeated on each chromosome. In some cases, these sequences are believed to be transposons, which are present in many copies in the human genome. After the crossover, the long chromosome has two copies of the globin gene. Independent mutations in each of the genes, which could arise in subsequent generations, would lead to their diversification. (Since the short chromosome lacks the original globin gene, the individuals that inherit this chromosome would be expected to be lost from the population.)

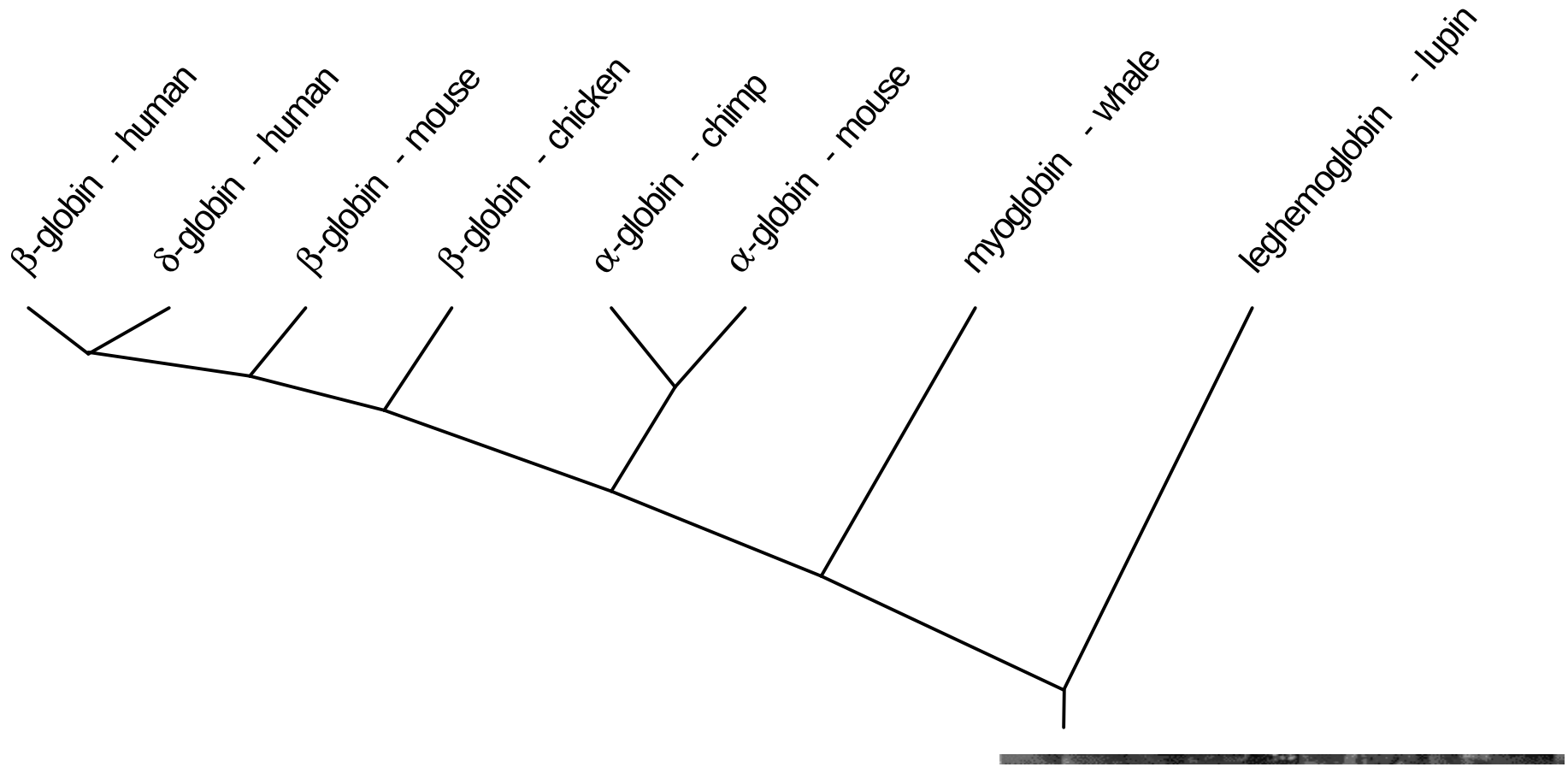
Similariteit vs. homologie

- Sequenties zijn *similair* als ze voldoende gelijkenis vertonen op sequentieniveau (DNA, proteïne, …)

- Similariteit kan voortvloeien uit
  - Homologie
  - Convergentie (functionele bepalingen)
  - Toeval

- Sequenties zijn *homoloog* als ze stammen uit een gemeenschappelijke voorouder
  - Homologe sequenties zijn *paraloog* als hun verschillen voortvloeien uit een genduplicatie
  - Homologe sequenties zijn *ortholoog* als hun verschillen niet gerelateerd zijn aan een genduplicatie

Orthologie vs. paralogie



Phylogenie

- Relaties tussen genen of proteïnen kunnen worden afgeleid op basis van hun sequentie
- Reconstructie van de moleculaire evolutie = phylogenie

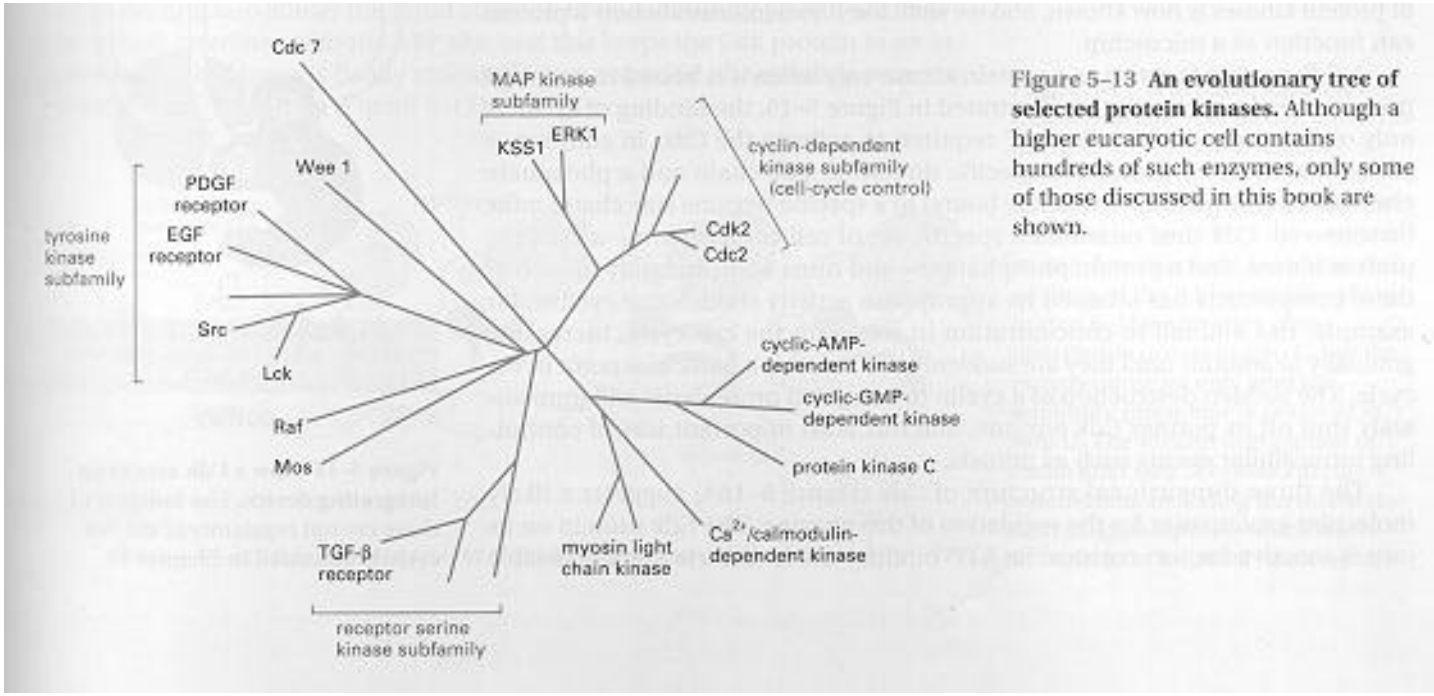


Figure 5–13 **An evolutionary tree of selected protein kinases.** Although a higher eucaryotic cell contains hundreds of such enzymes, only some of those discussed in this book are shown.

Homologie voor structuur- en functievoorspelling

- Homologe proteïnen hebben gelijkaardige structuren
- Homologe proteïnen hebben potentieel gelijkaardige functies (orthologen: cellulaire rol; paralogen: biochemische functie)

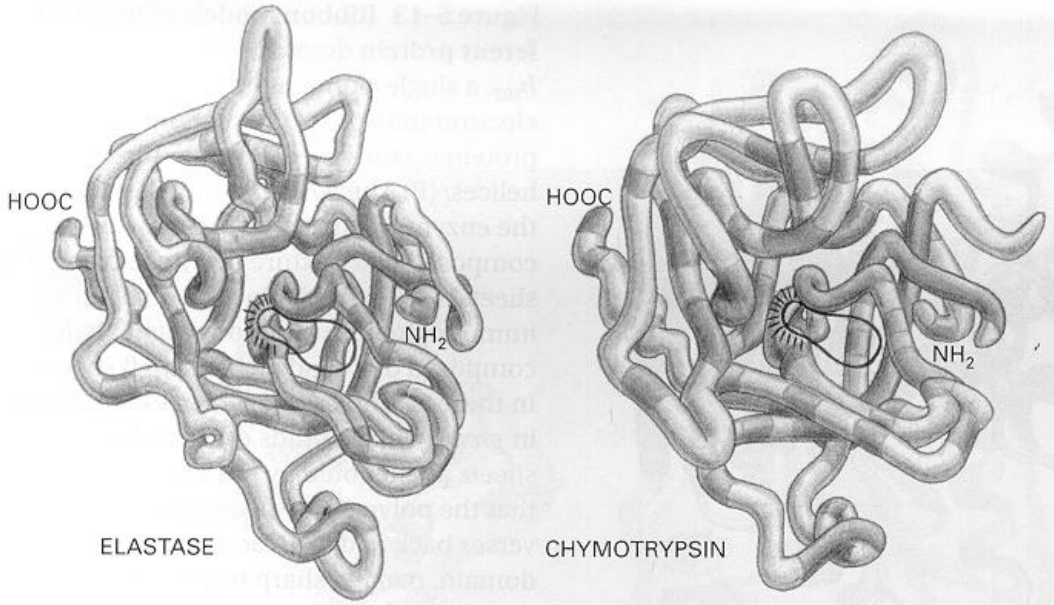


Figure 5–14 Comparison of the conformations of two serine proteases. The backbone conformations of elastase and chymotrypsin. Although only those amino acids in the polypeptide chain shaded in *green* are the same in the two proteins, the two conformations are very similar nearly everywhere. The active site of each enzyme is circled in *red;* this is where the peptide bonds of the proteins that serve as substrates are bound and cleaved by hydrolysis. The serine proteases derive their name from the amino acid serine, whose side chain is part of the active site of each enzyme and directly participates in the cleavage reaction.

Homologie voor voorspelling op het DNA

- Geconserveerde gebieden vloeien voort uit selectiedruk en zijn dus functioneel belangrijk
  - Genen
  - Regulatiegebieden
- Comparatieve genomica
- Genen kunnen voorspeld worden door de vergelijking van genomen op geschikte evolutieafstand (bvb., muis en mens)

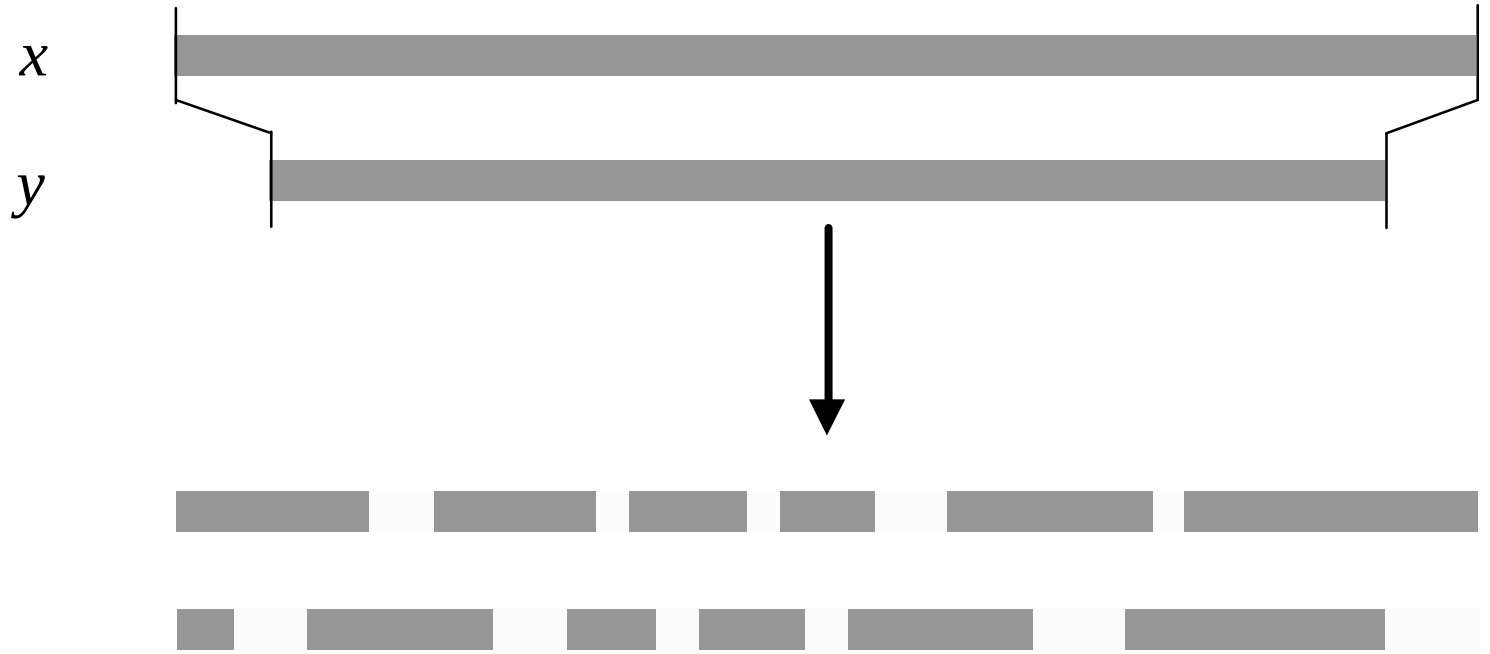
Principes van alignering

Elementen van een alignering

- Type van alignering
  - DNA vs. proteïne
  - Paarsgewijs vs. meervoudige alignering
  - Globale alignering
  - Lokale alignering
- Scoringsmodel voor aligneringen
  - Substituties
  - Gaps (inserties, deleties)
  - Substitutiematrix en gapkost
- Algoritme
  - Dynamisch programmeren
  - Heuristiek
- Significantieberekening

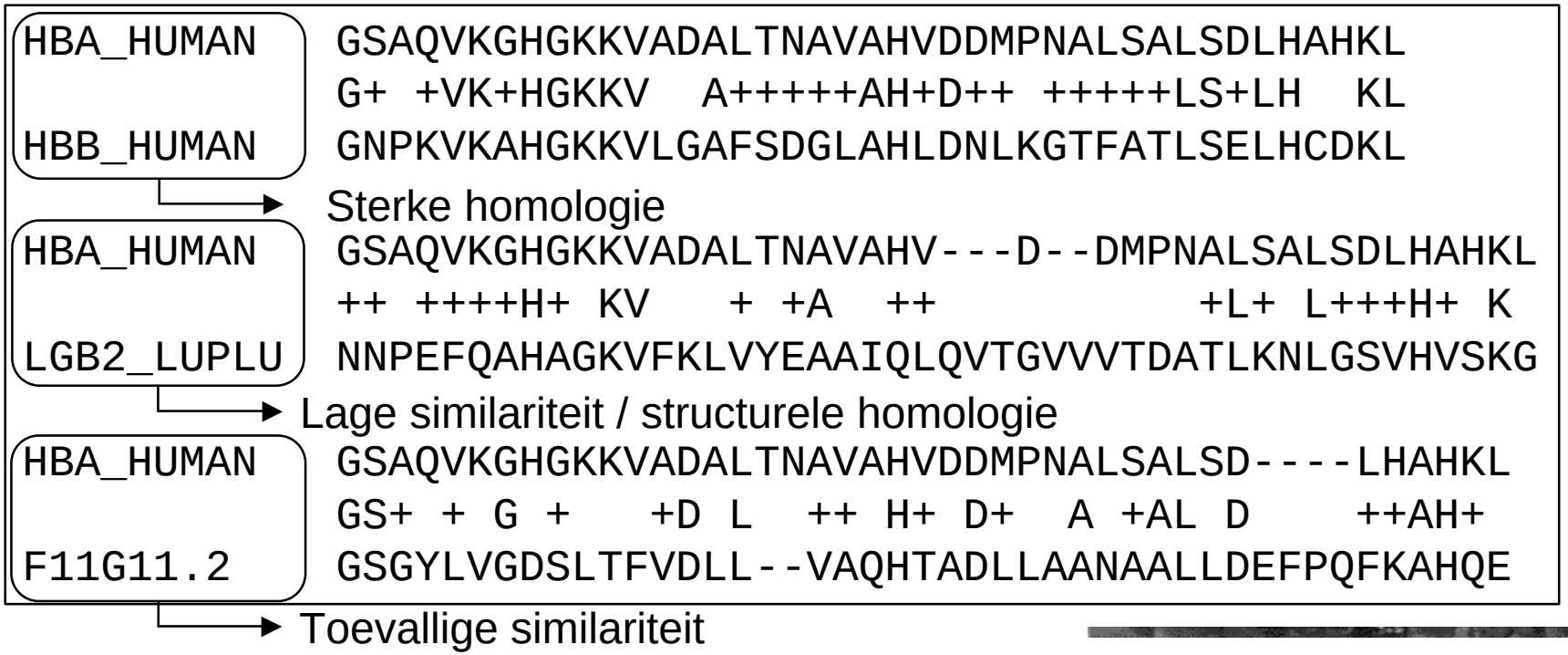
```
HEAGAWGHE-E
--P-AW-HEAE
```

Globale aligning

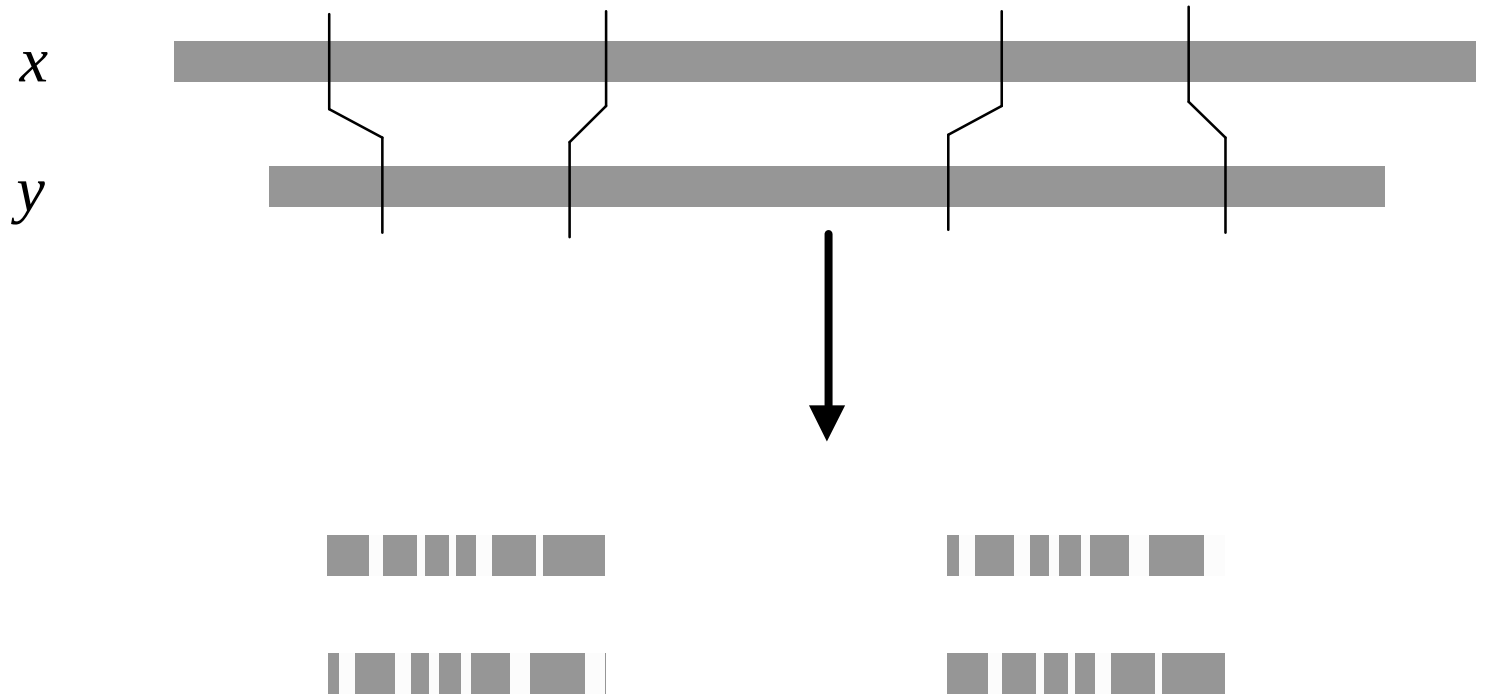


Globale aignering

- Aignering van ‘human alpha globin’ tegen ‘human beta globin’, ‘lupin leghemoglobin’ en ‘glutathionine S-transferase homolog F11G11.2’ (‘+’ voor goede substituties)



Lokale aignering



Substitutiematrix en gapkost

- Het aligneren van twee residus kan meer of minder waarschijnlijk zijn
- Om de kwaliteit van een alignering te berekenen, brengt het aligneren van twee residues een winst of kost met zich mee
- Substitutiematrix
  = BLOSUM50
- Gaps hebben ook een kost

```
HEAGAWGHE-E
--P-AW-HEAE
```

BLOSUM50 substitutiematrix

	A	R	N	D	C	Q	EGHILKMFPSTWYV
A	5	-2	-1	-2	-1	-1	.....
R	-2	7	-1	-2	-4	1	.....
N	-1	-1	7	2	-2	0	.....
D	-2	-2	2	8	-4	0	.....
C	-1	-4	-2	-4	13	-3	.....
Q	-1	1	0	0	-3	7	.....
...	...	...	...	...	...	...	.....

Substitutiematrix voor DNA

- Standard

	A	C	G	T
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5

Dynamisch programmeren

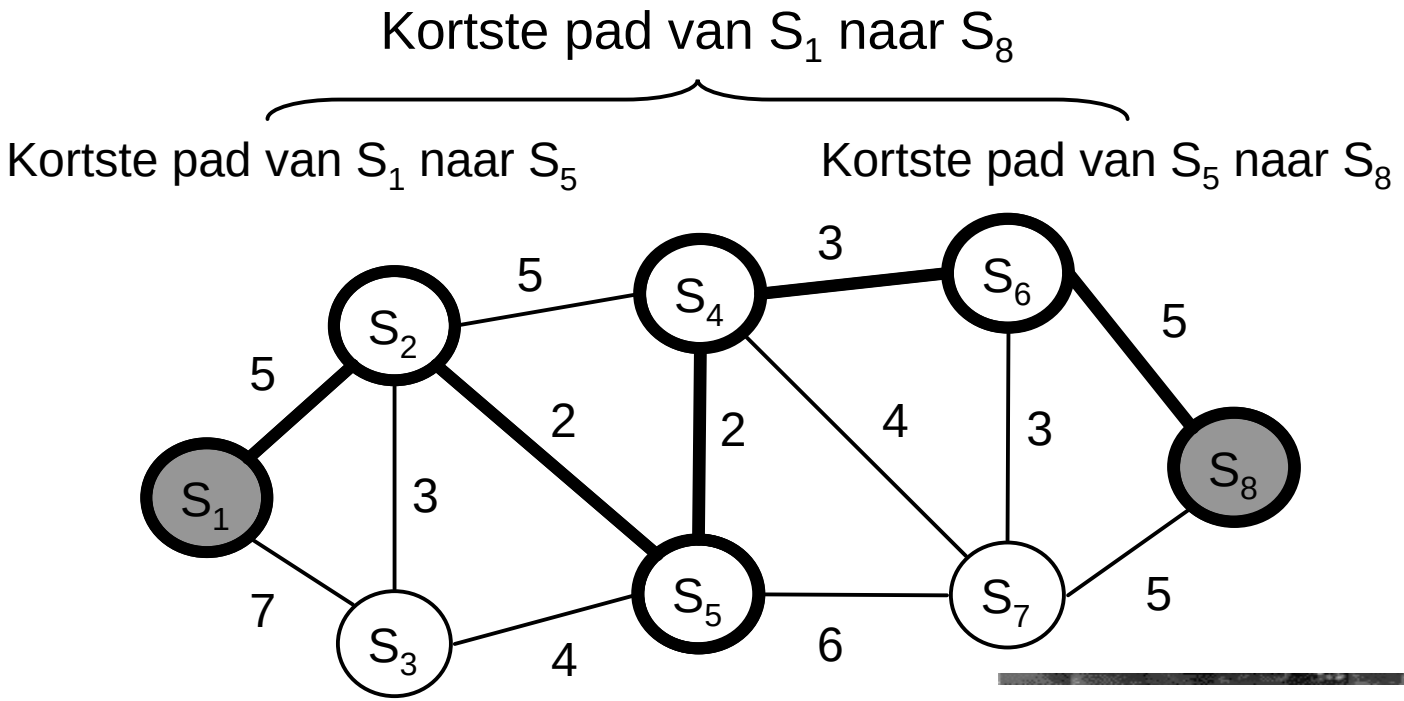
- Aligneren is het vinden van het minimum-kost/maximum-winst pad door de kosttabel = DYNAMISCH PROGRAMMEREN
- Gapkost = -8

	*	H	E	A	G	A	W	G	H	E	E
*	0	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8
P	-8	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-8	-2	-1	5	0	5	-3	0	-2	-1	-1
W	-8	-3	-3	-3	-3	-3	15	-3	-3	-3	-3
H	-8	10	0	-2	-2	-2	-3	-2	10	0	0
E	-8	0	6	-1	-3	-1	-3	-3	0	6	6
A	-8	-2	-1	5	0	5	-3	0	-2	-1	-1
E	-8	0	6	-1	-3	-1	-3	-3	0	6	6

```
HEAGAWGHE-E
--P-AW-HEAE
```

Dynamisch programmeren

- Optimaliteitsprincipe van Belman
- Voorbeeld : vinden van de kortste treinroute tussen twee steden



Aligneringsalgoritmen

Globale aligning

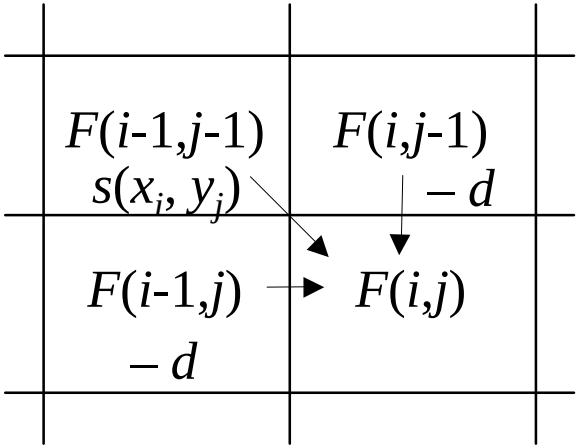
- Needleman-Wunsch algoritme
- Progressief invullen van een tabel $F(i,j)$ (!!! kolom, rij) die de maximum winst bijhoudt voor de alignering van sequentie x tot x_i met sequentie y tot y_j

  I G A x_i — A I G A x_i — G A x_i - -

  L G V y_j — G V y_j - - — S L G V y_j
- Substitutie matrix $s(x, y)$ en gapkost d
- Recurrentie

$$F(i,j) = \max \begin{cases} F(i\text{-}1,j\text{-}1) + s(x_i, y_j) & \text{substitutie} \\ F(i\text{-}1,j) - d & \text{deletie} \\ F(i,j\text{-}1) - d & \text{insertie} \end{cases}$$

$$F(0,0) = 0$$



- Start links boven
- Vul progressief in met recurrentie
- Gebruik traceback pointers

$$F(i,j) = \max \begin{cases} F(i\text{-}1,j\text{-}1) + s(x_i, y_j) \\ F(i\text{-}1,j) - d \\ F(i,j\text{-}1) - d \end{cases}$$

	*	H	E	A	G	A	W	G	H	E	E
*	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

Lokale alignering

- Smith-Waterman algoritme
- Beste alignering tussen subsequenties van x en y
- Als de huidige alignering een negatieve score heeft, is het beter een nieuwe alignering te starten

$$
F(i,j) = \max \begin{cases} 0 & \text{restart} \\ F(i\text{-}1,j\text{-}1) + s(x_i, y_j) & \text{substitutie} \\ F(i\text{-}1,j) - d & \text{deletie} \\ F(i,j\text{-}1) - d & \text{insertie} \end{cases}
$$

$$
F(0,0) = 0
$$

- Start links boven
- Vul progressief in met recurrentie
- Traceback vanaf de hoogste score en stop op nul

	*	H	E	A	G	A	W	G	H	E	E
*	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0	0
W	0	0	0	0	2	0	20	12	4	0	0
H	0	10	2	0	0	0	12	18	22	14	6
E	0	2	16	8	0	0	4	10	18	28	-19
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

AWGHE
AW-HE

Aligneren met affiene gapkost

- Lineaire kost voor de gap is vaak te streng
- Willekeurige gapkost verhoogt complexiteit van $O(n^2)$ naar $O(n^3)$ (lookbacks op grote afstand)
- $O(n^2)$ complexiteit voor affiene gapkost = $-d-(g-1)e$
- Het is dan nodig om drie variabelen M, I_x, I_y (match, deletie, insertie) te tracken in plaats van enkel F

I G A x_i
L G V y_j

A I G A x_i
G V y_j - -

G A x_i - -
S L G V y_j

M: max score | $x_i \# y_j$ I_x : max score | $x_i \#$ - I_y : max score | $y_j \#$ -

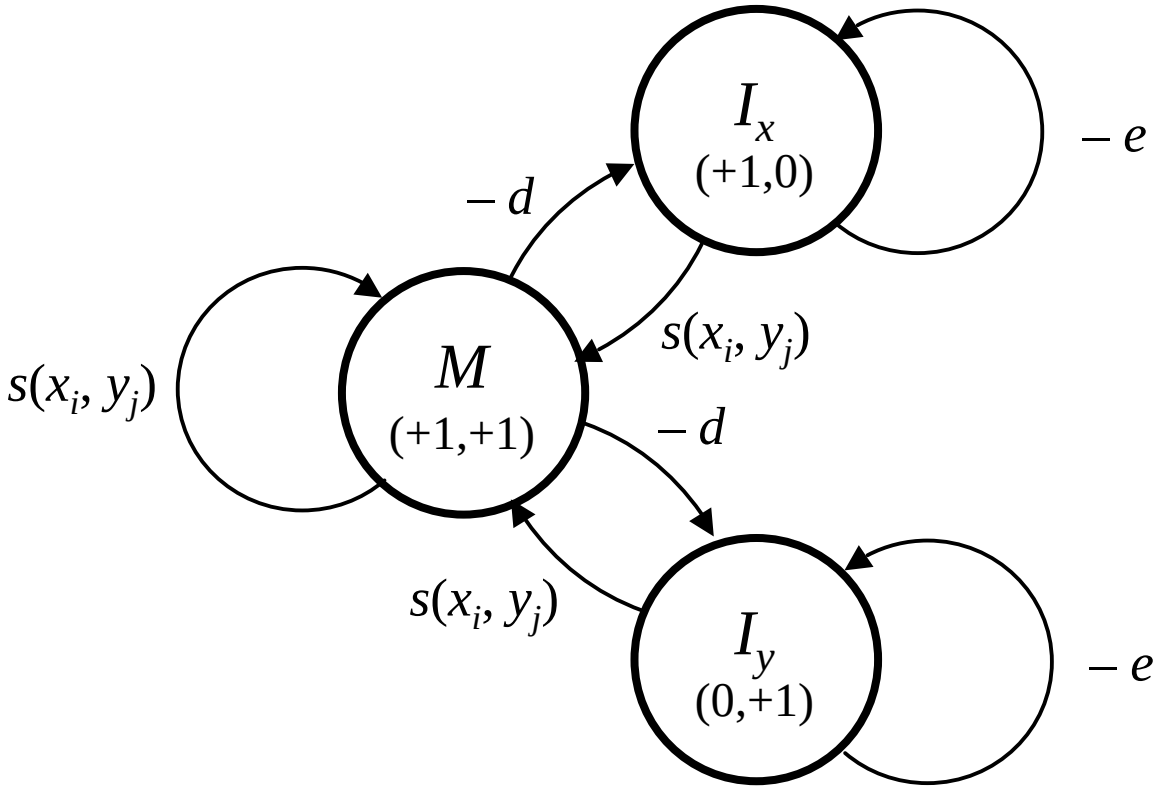
Recurrentie

- $-d-e$ kleiner dan kleinste substitutiekost om te vermijden dat een deletie direct door een insertie gevolgd wordt
- Alignering gevonden door het maximum over de drie variabelen te tracken

$$
\begin{aligned}
M(i,j) &= \max \begin{cases} M(i\text{-}1,j\text{-}1) + s(x_i, y_j) \\ I_x(i\text{-}1,j\text{-}1) + s(x_i, y_j) \\ I_y(i\text{-}1,j\text{-}1) + s(x_i, y_j) \end{cases} \\
I_x(i,j) &= \max \begin{cases} M(i\text{-}1,j\text{-}1) - d \\ I_x(i\text{-}1,j\text{-}1) - e \end{cases} \\
I_y(i,j) &= \max \begin{cases} M(i\text{-}1,j\text{-}1) - d \\ I_y(i\text{-}1,j\text{-}1) - e \end{cases}
\end{aligned}
$$

Finite state automaton

- Voorstelling van het algoritme als een finite state automaton



Significantieberekening

- Laten we kijken naar de *distributie* van N aligneringscores S t.o.v. random sequenties
- Voor een ungapped alignering is de score van een match de som van veel gelijkaardige random bijdragen en volgt daarom een normale distributie
- Voor een normale distributie volgt de distributie van het maximum M_N van een reeks van N random variabelen de *extreme value distribution* (EVD)

$$P(M_N <= x) = \exp(-KNe^{\lambda(x-\mu)})$$

Significantieberekening

- Voor ungapped aligneringen heeft (hoewel de random variabelen niet normaal verdeeld zijn) de EVD de vorm

$$P(S<=x) = \exp(Kmne^{-\lambda S})$$

met n lengte van de query, m lengte van de databank

- Verwachte score van een random sequentie met probabiliteit P_i voor de aminozuren moet negatief zijn :

$$-\sum_{i,j} P_i P_j s(i,j) < 0$$

- Ungapped alignering: parameters afgeleid uit P_i and $s(i,j)$
- Gapped aligneringen: parameters geschat via regressie
- Een alignering is relevant als de probabiliteit voldoende laag is (bvb., $P<0.01$)

Normalizatie van scores

- Verschillende substitutiematrices geven verschillende scaleringen van de scores

- Met P_i and $s(i,j)$ kan de score S van een ungapped alignering genormalizeerd worden via de twee afgeleide parameters λ and K :

$$S' = (\lambda S - \ln K)/\ln 2$$ (uitgedrukt in bits)

Lengtecorrectie

- Als de sequenties in de databank van verschillende lengte zijn, zullen de langere sequenties hogere scores krijgen in het algemeen

- Dit kan verbeterd worden door de gemiddelde score als functie van het logaritme van de lengte te berekenen en een lineaire functie hierdoor te fitten en af te trekken

Substitutiematrices

- Hoe een redelijke substitutiematrix kiezen?
- Kijken naar een verzameling van bevestigde aligneringen (met gaps) en de aminozuurfrequenties q_a, de substitutiefrequenties p_{ab}, en de gapfunctie $f(g)$ berekenen
- Aannemelijkheidsmodel
  - Random sequenties : $P(x,y|\mathrm{R}) = \Pi_i q_{xi} \Pi_j q_{yj}$
  - Alignering : $P(x,y|\mathrm{M}) = \Pi_i p_{xixj}$
  - Odds ratios : $P(x,y|\mathrm{M})/P(x,y|\mathrm{R}) = \Pi_i p_{xixj}/(\Pi_i q_{xi} \Pi_j q_{yj})$
  - Log-odds score : $S(x,y) = \Sigma_i s(x_i,y_i)$ with $s(a,b) = \log(p_{ab}/q_a q_b)$
- Substitutiematrix $s(a,b) = \log(p_{ab}/q_a q_b)$

PAM matrix

- Point Accepted Mutation matrix
- Problemen
  - Aligneringen zijn niet onafhankelijk voor gerelateerde proteïnen
  - Verschillende aligneringen komen overeen met verschillende evolutietijden
- PAM1 matrix
  - Boom van proteïnefamilies
  - Schat ancestrale sequenties
  - Schat mutaties op korte evolutieafstand
  - Scalering naar een substitutiematrix
  - 1% Point Accepted Mutation (PAM1)
- PAM250 is 250% Point Accepted Mutations (~20% similariteit) = 250ste macht van PAM1

BLOSUM matrix

- BLOCKS SUbstitution Matrix
- PAM werkt minder goed op grotere evolutieafstanden
- Ungapped aligneringen van proteïnefamilies uit de BLOCKS databank
- Groepering van de sequenties met meer dan $L\%$ identieke aminozuren (bvb., BLOSUM62)
- Substitutiefrequentie van aminozuren tussen de verschillende groepen (met verbetering i.f.v. de groepsgrootte) wordt gebruikt om de substitutiematrix af te leiden

BLAST

BLAST

- Voor heel grote databanken is Smith-Waterman lokale alignering te traag
- Basic Local Alignment Search Tool (BLAST) is een snel heuristiek algoritme voor lokale alignering (http://www.ncbi.nlm.nih.gov/Entrez)
  - BLASTP – proteïnequery op proteïnedatabank
  - BLASTN – nucleotidequery op nucleotidedatabank
  - BLASTX – vertaalde nucleotidequery op proteïnedatabank (vertaling naar de zes leesramen)
  - TBLASTN – proteïnequery op vertaalde nucleotidedatabank
  - TBLASTX – vertaalde nucleotidequery op vertaalde nucleotidedb

BLASTP

- Stap 1 : Vind alle woorden van lengte w (bvb., w=3) waarvoor er een match is in de querysequentie met score groter of gelijk aan T (bvb., T=11) met de gekozen substitutiematrix (bvb., BLOSUM62 met gapkost $10+g$)

- Stap 2 : Gebruik een finite state automaton om alle matches met de woordlijst te vinden in de databank (*hits*)

BLASTP

- Stap 3 : Check voor welke hits er een andere hit zonder overlap bestaat binnen een afstand van A (bvb., A=40) (de afstand moet dezelfde zijn op de query als op de target) (*two-hits*)

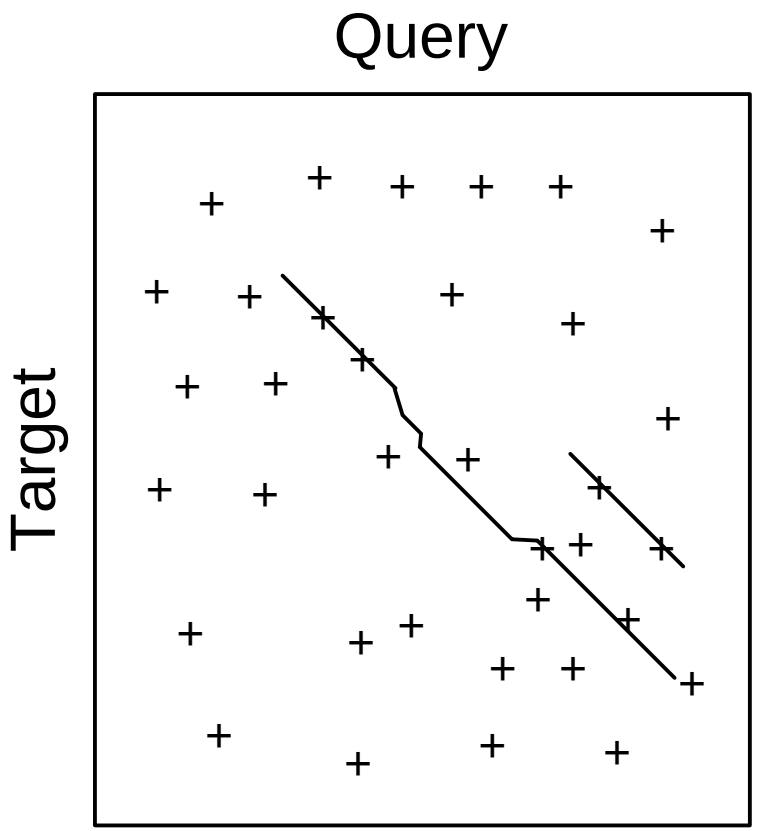
- Stap 4 : Breid de linkse hit van de two-hits uit in beide richtingen via ungapped alignering ; stop de uitbreiding wanneer de score met X_g (bvb., X_g=40) valt onder de beste score dusver (*high scoring segment pair* HSP)

BLASTP

- Stap 5 : Breid de HSPs met genormalizeerde score boven S_g (S_g =22 bits) via gapped alignering en stop de uitbreiding wanneer de score met X_g (bvb., X_g=40) valt onder de beste score tot nu toe ; selecteer de beste gapped lokale alignering

- Stap 6 : Bereken de significantie van de aligneringen ; voor de significante aligneringen herhaal de gapped aligneringen met een hogere dropoff parameter X_g voor meer nauwkeurigheid

BLASTP



Hits

Two-hits

Lokale aignering

BLASTP example

Protein family	Query (SWISS-PROT)	Smith-Waterman	BLAST (# matches)
Serine protease	P00762	275	275
Serine protease inhibitor	P01008	108	108
Ras	P01111	255	252
Globin	P02232	28	28
Hemagglutinin	P03435	128	128
Interferon alpha	P05013	53	53
Alcohol dehydrogenase	P07327	138	137
Histocompatibility antigen	P10318	262	261
Cytochrome P450	P10635	211	211
Glutathione transferase	P14942	83	81
H^+-transport ATP synthase	P20705	198	197
Running time		36	0.34

Samenvatting

- Similariteit vs. homologie
- Dynamisch programmeren
- Lokale alignering
  - Needleman – Wunsch
- Globale alignering
  - Smith – Waterman
  - Affiene gapkost
- Substitutiematrices
  - PAM
  - BLOSUM
- Significantieberekening
- BLAST