

HK07 – Les 3
Meervoudige alignering

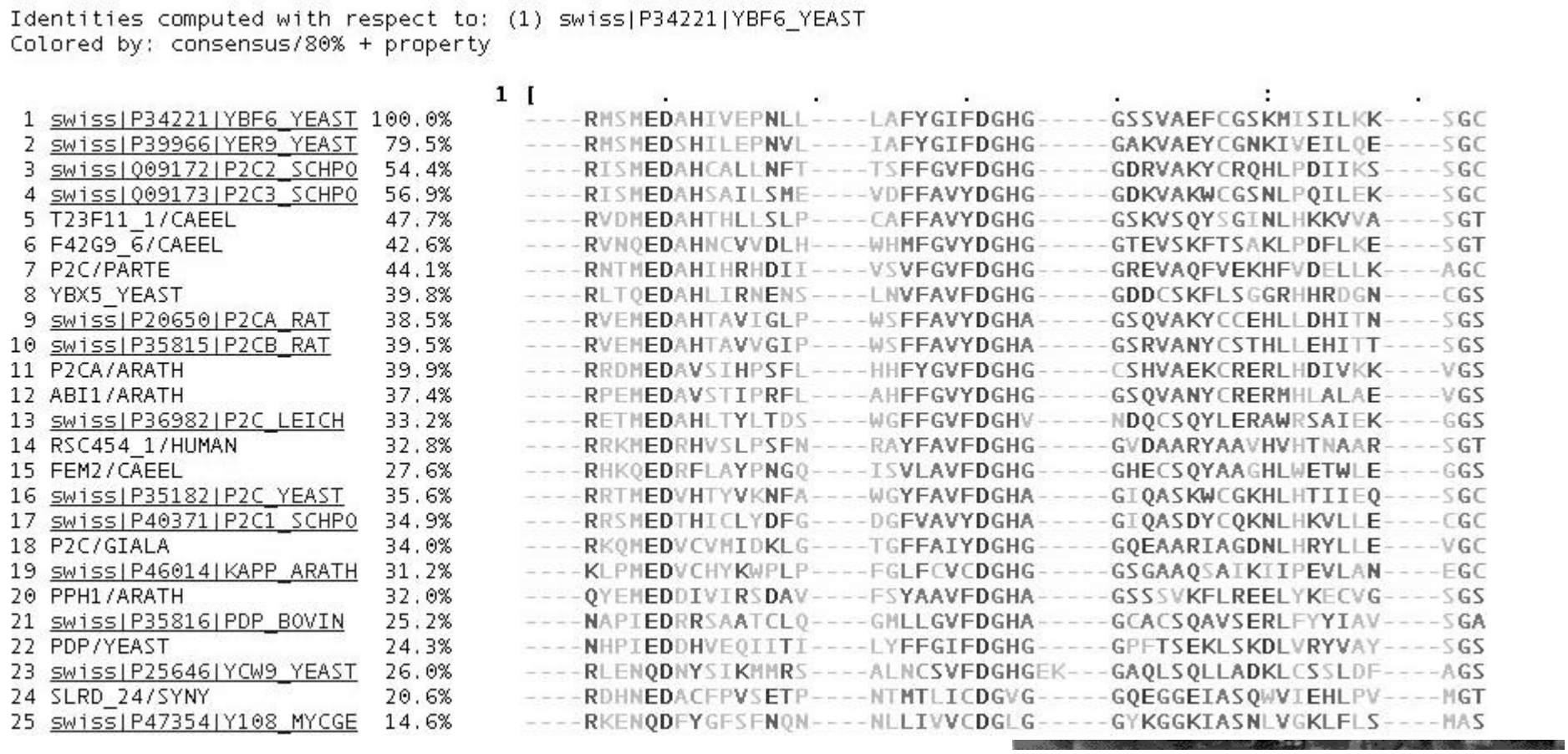
Yves Moreau
3de jr. Burg. Ir. Elektrotechniek
Dataverwerking & Automatisatie
2001-2002

Meervoudige alignering

- Hoe kunnen we de sequenties van meerdere proteïnen uit dezelfde familie tegelijkertijd vergelijken (dus niet paarsgewijs) ?
- Meervoudige alignering zijn bvb. belangrijk om de rol van de residus te onderzoeken in de 3D structuur van de proteïnen (idealiter, zouden gealigneerde residus dezelfde 3D coordinate moeten hebben en zouden evolutionair gerelateerd moeten zijn)

- Handmatig opbouwen van meervoudige aligneringen is tijdsrovend

Meervoudige aligning



Meervoudige alignering

- Elementen van een meervoudige alignering
  - Scoringsmodel
  - Algoritme om de beste alignering te vinden

- In theorie (gegeven de fylogenetische boom van de sequenties) is de waarschijnlijkheid van een meervoudige alignering het product van de waarschijnlijkheden van alle evenementen die nodig zijn om deze boom te produceren (in praktijk hebben we nooit genoeg gegevens om dit uit te rekenen)

Scoringsmodel

- Een pragmatische scoringsfunctie heeft de vorm

$$S(m) = G + \Sigma_i S(m_i)$$

  G gapkost, m_i kolom i van de meervoudige alignering, $S(m_i)$ score voor kolom i

- Kolomscore: sum-of-pairs score

$$S(m_i)=\Sigma_l \Sigma_{k<l} s(m_i^k, m_i^l)$$

  $s()$ substitutiematrix

- Sum-of-pairs score is niet statistisch onderbouwd
  - Uitbreiding van log-odds score (3 seqs.): $\log(p_{abc}/(q_a q_b q_c))$
  - Sum of pairs: $\log(p_{ab}/(q_a q_b)) + \log(p_{bc}/(q_b q_c)) + \log(p_{ac}/(q_a q_c))$

PSI-BLAST

PSI-BLAST

- Position-Specific Iterated BLAST
- Het zoeken in een databank met een scorematrix die een proteïnefamilie beschrijft is veel gevoeliger dan het gebruik van een enkelvoudige querysequentie
- Idee:
  - Run BLAST op de querysequentie
  - Verzamel de significante aligneringen
  - Bouw een specifieke matrix die de gealigneerde sequenties beschrijft
  - Run BLAST met deze specifieke matrix om meer sequenties te rekruteren
  - Bouw een nieuwe matrix en itereer tot er geen nieuwe sequentie meer worden gerekruteerd

PSI-BLAST

- Motieven zijn aanwezig in proteïnefamilies ; indien wij zulke motieven gebruiken in onze zoekbewerkingen, kunnen we verdere familieleden opsporen
- Het is moeilijk om meerdere motieven automatisch te bepalen in een reeks sequenties ; daarom bouwt PSI-BLAST een enkelvoudige motief die even land is als de query (Lx20 matrix)
- De motief is *geen* substitutiematrix maar geeft *voor elke positie* een specifieke score voor de aanwezigheid van ieder aminozuur en wordt daarom *positiespecifieke scoringsmatrix* (PSSM) genoemd (gapkost is niet positiespecifiek)

Positiespecifieke scoringsmatrix

```
..............MFGKRAFVHHYVGEGMEENEFTDARQDLYELEVDYANL..............
..............MFKRKGFLHWYTGEGMEPVEFSEAQSDLEDLILEYQQY..............
..............MFSRKAFLHWYTGEGMEEGDFAEADNNVSDLLSEYQQY..............
```

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z	*	-
	-56	-57	-56	-58	3	-54	-61	-51	-48	-47	-46	35	-56	-62	10	22	-58	-55	-51	-58	-42	-58	7	0	0
.....	2	7	-36	-1	0	24	-9	-32	-38	-40	-38	-38	18	-47	-38	-41	0	-37	-39	-32	-19	19	-21	0	0
	20	-46	-39	-47	-44	10	-8	-40	-45	8	-46	-43	-44	-48	-41	16	-5	-41	-18	-44	-22	11	-42	0	0
.....	-44	-45	-44	-46	-42	20	-48	-38	-41	19	0	-38	-42	-50	-37	22	-45	-44	-43	-41	-28	4	-39	0	0
	-1	-6	-41	6	-36	-40	-43	-37	-36	20	8	14	-23	-45	-19	17	-40	-40	-38	-47	-21	-43	-27	0	0
.....	21	8	-44	3	14	6	-14	-45	-51	-46	-52	-50	13	-4	-44	-26	-42	-45	-47	-53	-25	-48	-19	0	0
	-48	-29	-44	-51	-48	29	-53	-42	8	-47	-41	-41	-2	-57	18	-49	1	-47	-42	-40	-31	13	-11	0	0
.....	-36	-16	-32	3	-40	17	-44	-34	-22	-37	12	14	-40	-46	-38	7	-40	-35	14	-31	-20	12	-39	0	0
	-50	-46	-53	-49	-45	8	-53	36	-54	17	-53	-49	-43	-55	-1	11	-49	-50	-54	-51	-37	-39	-21	0	0
.....	-42	-45	-42	-47	-43	16	-46	5	-44	11	-43	-40	-43	-50	2	3	1	-43	-44	42	-28	-27	-18	0	0
	-50	-48	-51	-46	18	-19	-56	-41	-53	-50	-54	-53	-51	7	-47	-52	1	-3	-53	-41	-33	32	-18	0	0
.....	9	-12	-33	7	-37	-40	-40	-40	-30	8	-35	16	-37	-44	-38	-37	-7	18	14	-45	-19	-43	-38	0	0
	-41	-4	-41	5	-46	10	22	-46	15	-47	-24	18	-15	-6	-47	-48	-14	-44	-35	-46	-22	-41	-47	0	0
.....	2	-20	-51	3	24	22	2	-47	-52	-48	-51	-51	-49	-56	-44	-51	-49	-51	-51	-50	-32	-41	-14	0	0
	-36	6	-43	7	6	-48	23	-36	-50	6	-49	-43	5	-46	15	-36	-37	-40	-47	-47	-25	-47	11	0	0
	-55	14	-54	18	-52	6	2	-54	-47	-6	-45	35	8	-63	-53	-55	-56	-54	-50	-57	-35	-53	-52	0	0
.....	-40	22	-44	24	17	-50	-9	-36	-50	-38	-52	-48	19	-47	-35	-41	0	-38	-48	-53	-29	-48	-12	0	0
	-52	-49	-52	-45	23	15	-59	-50	-18	-51	16	-42	-54	-10	-46	-53	-54	-53	-47	-53	-34	-46	-16	0	0
.....	-35	13	-37	18	3	-40	13	-36	-35	-37	-20	24	8	-44	-35	-39	-1	4	-11	-44	-17	-43	-19	0	0
	-73	-37	-82	-10	31	-82	-78	-69	-78	-68	-78	13	-71	-78	-62	-72	-72	-74	-76	-82	-62	-81	-21	0	0
.....	-63	-70	-55	-71	-71	34	-69	-59	-50	-66	1	24	-69	-71	-65	-66	-66	-62	-54	-50	-51	-40	-68	0	0
	-3	-2	-41	25	3	-50	-42	-42	-48	-42	-51	-47	-35	-48	-42	-45	10	17	-44	-54	-28	-49	-22	0	0
.....	-51	3	-57	8	27	-62	-56	-49	-5	-49	-61	-57	-4	-58	-45	-53	9	-6	-55	-64	-34	-60	-13	0	0
	25	-57	-48	-58	-56	-61	-51	-57	-58	-56	-60	-58	-55	-58	-56	-58	18	-49	-17	-64	-42	-61	-56	0	0
.....	-49	-18	-55	3	14	-55	-53	-43	-53	4	-52	13	-46	-55	-8	28	-49	-50	-52	-56	-34	-54	2	0	0
	0	3	-44	11	22	-50	-23	13	-50	-39	-51	-47	-7	-47	-2	-41	13	-38	-46	-52	-24	-46	9	0	0
	-44	18	-40	17	-41	-44	-46	-41	20	-45	-39	-38	20	-52	-45	-46	-42	7	5	-52	-28	-47	-43	0	0
	-41	-51	-36	-53	-50	-37	-52	-48	4	-48	9	27	-50	-52	-46	-48	-45	5	21	-45	-30	-44	-47	0	0
	11	10	-39	3	3	-43	-39	-32	-44	-3	-22	3	18	-45	23	-36	-12	-37	-41	-44	-18	1	14	0	0
	10	17	16	24	-5	-47	-38	-37	-45	-39	-48	-44	7	-45	6	-41	8	-36	-20	-50	-23	-46	1	0	0

Constructie van de meervoudige alignering

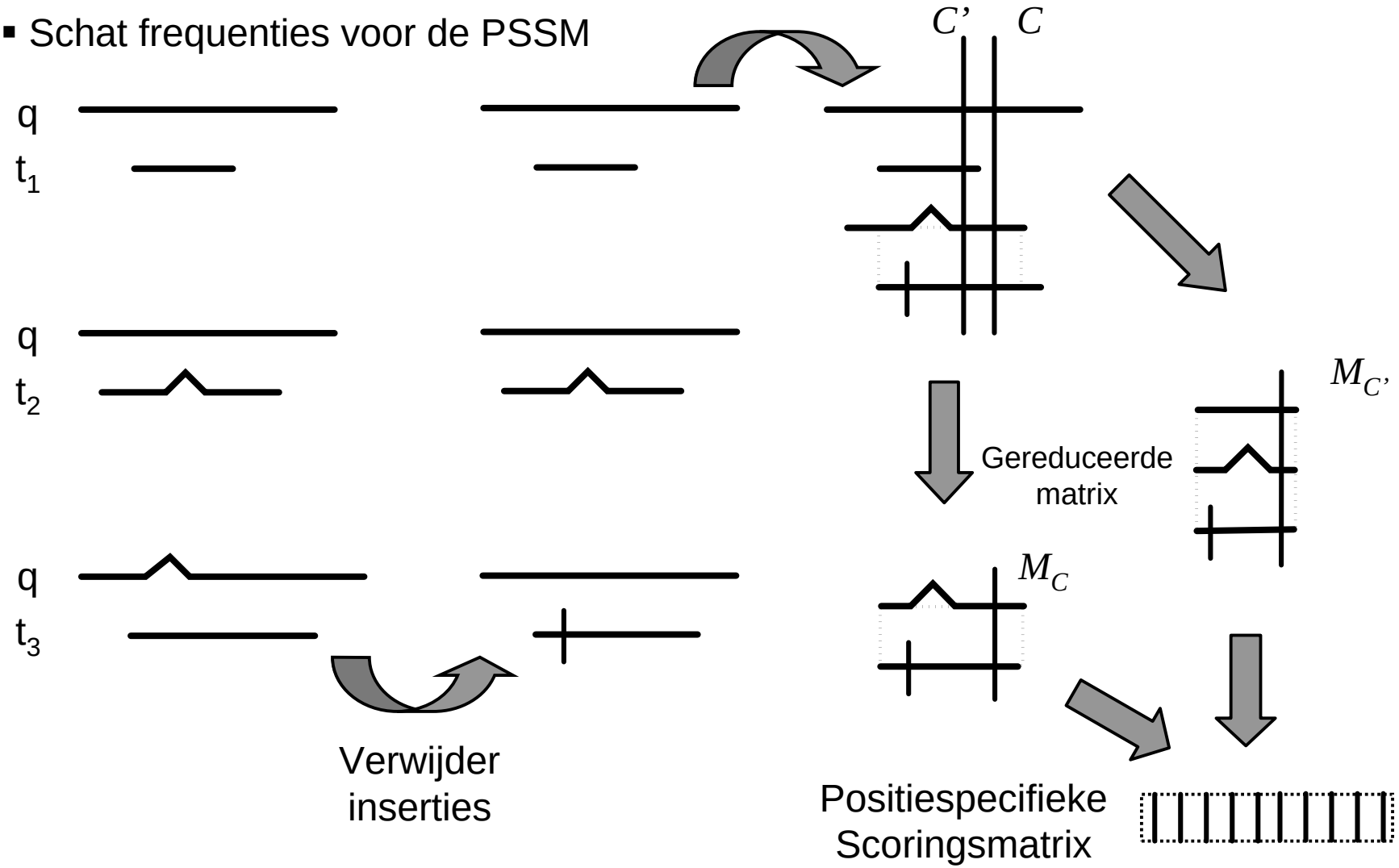
- BLAST de querysequentie
- Verzamel alle gealigneerde sequentie met P een bepaalde drempel (bvb., $P<0.01$)
- Verwijder alle sequenties die identiek zijn aan de query
- Bewaar enkel een instantie van alle sequenties die meer dan 98% identiek zijn met elkaar
- Schat de frequentie van ieder aminozuur op iedere positie (*MOEIJLIJK!!!*)

Schatting van de frequenties

- Om de ruis te verminderen in de schatting van de frequenties, zijn de matrixscores voor een bepaalde positie niet enkel afhankelijk van de betrokken kolom maar ook van andere kolommen
- Aangezien we lokale alignering gebruiken, dekken de gealigneerde sequenties niet de volledige query – we moeten daarom de gegevensmatrix reduceren om de berekeningen mogelijk te maken
  - Voor de aligneringen met een insertie in de query worden de betrokken kolommen verwijderd
  - Voor iedere kolom C, bouw de gereduceerde M_C door enkel de sequenties te beschouwen die een residu op C hebben en door enkel de kolommen te beschouwen waar al deze sequenties aanwezig zijn
- Schat de frequenties met behulp van de gereduceerde matrix

Contextafhankelijke matrixscore

- Vind gealigneerde matches
- Verwijder inserties
- Voor kolom C, bouw de gereduceerde matrix M_C
- Idem voor C'
- Schat frequenties voor de PSSM



Schatting van de frequenties

- Aangezien gelijkaardige sequenties redundant zijn kan hun bijdrage tot de scorematrix best gesplitst worden ; dus we gebruiken gewogen residufrequenties f_i voor residu $b=i$

- Het aantal onafhankelijke observaties voor een bepaalde kolom kan sterk varieren en is verschillend van het aantal sequenties ; tel het gemiddelde aantal residus N_C in de kolommen van M_C

- Schat de gegevensafhankelijke pseudocounts g_i en schat de frequenties met behulp van pseudocountverbeteringen

$$Q_i = (\alpha f_i + \beta g_i) / (\alpha + \beta)$$

(bvb., $\alpha = N_C - 1$, $\beta = 10$)

- Bereken de scores als log-odds ratios

$$s(i,C) = \log(Q_i/P_i)$$

met P_i achtergrondfrequenties
uit de substitutiematrix

BLAST met PSSMs

- Enkel beperkte aanpassingen zijn nodig om BLAST te runnen met een positiespecifieke scoringsmatrix (in essentie, vervangen we de score die we normaal gezien via de query en de substitutiematrix $s(t_i,q_i)$ bekomen door de score die we via de PSSM $s(t_i,C)$ bekomen)

- Om de parameters van BLAST niet te moeten modificeren, moeten de scores op een aangepaste manier herschaald worden

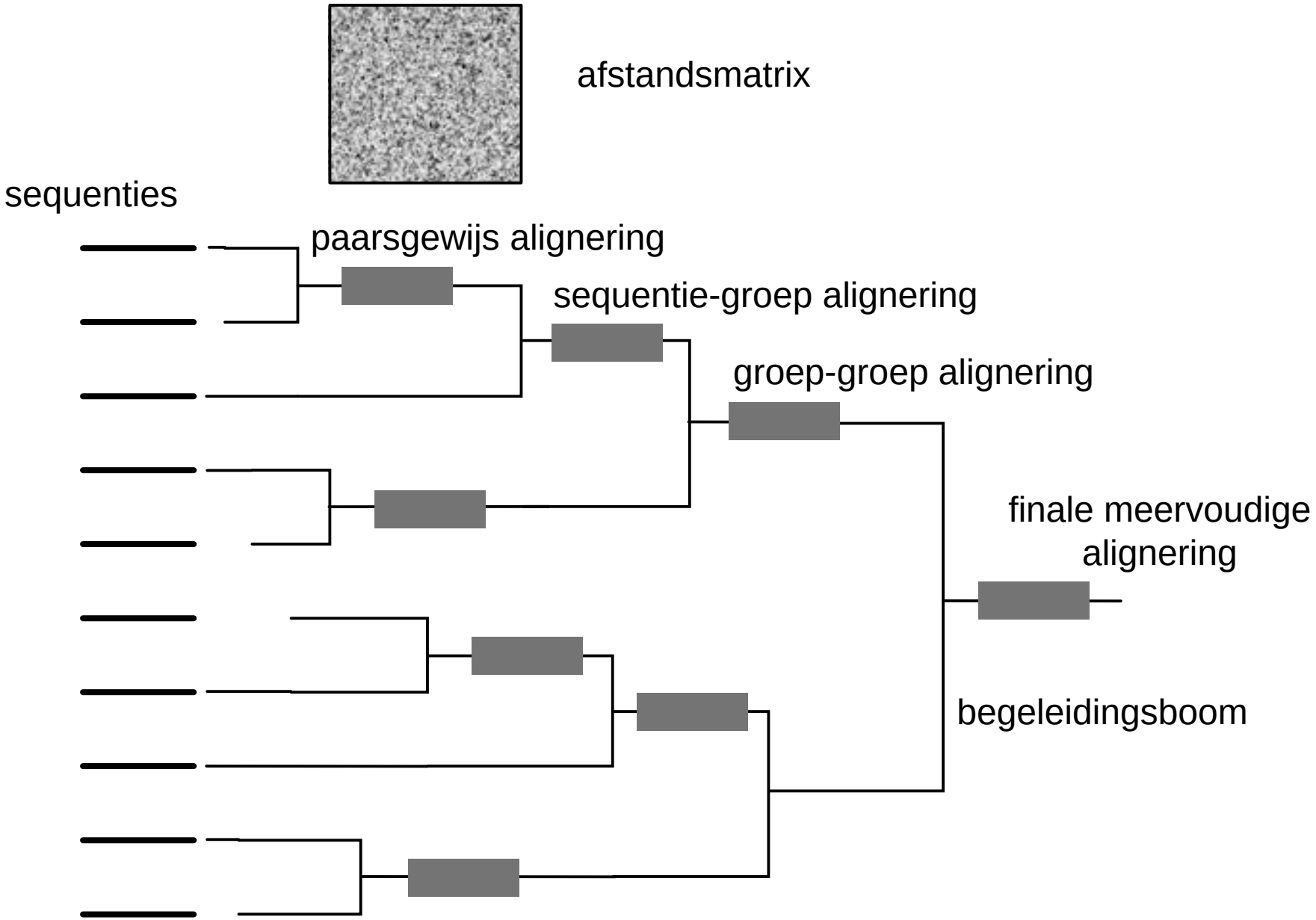
PSI-BLAST voorbeeld

Protein family	Query (SWISS-PROT)	Smith-Waterman	BLAST	PSI-BLAST (# matches)
Serine protease	P00762	275	275	286
Serine protease inhibitor	P01008	108	108	111
Ras	P01111	255	252	375
Globin	P02232	28	28	623
Hemagglutinin	P03435	128	128	130
Interferon alpha	P05013	53	53	53
Alcohol dehydrogenase	P07327	138	137	160
Histocompatibility antigen	P10318	262	261	338
Cytochrome P450	P10635	211	211	224
Glutathione transferase	P14942	83	81	142
H^+-transport ATP synthase	P20705	198	197	207
Running time		36	0.34	0.87

CLUSTALW

CLUSTALW

- Meervoudige alignering kan geformuleerd worden als multidimensioneel dynamisch programmeren, maar voor N sequenties van lengte L is de complexiteit $O(L^N)$ (onmogelijk in de praktijk dus)
- CLUSTALW (Cluster and ALign) is een voorbeeld van progressieve meervoudige alignering
  - Bouw een begeleidingsboom gebaseerd op sequentiesimilariteit (geen echte fylogenetische boom)
  - Doe paarsgewijs alignering langs de begeleidingsboom
    - Paasgewijs sequentiealignering
    - Sequentie–groep alignering
    - Groep–groep alignering


afstandsmatrix
sequenties
paarsgewijs alignering
sequentie-groep alignering
groep-groep alignering
finale meervoudige
alignering
begeleidingsboom

CLUSTALW

- Algoritme
  - Bouw afstandsmatrix voor alle sequentieparen
    - Bereken paarsgewijs ‘gapped’ aligneringen voor alle paren
    - Score met percentage verschillen tussen gealigneerde sequenties (gaps geexcludeerd)
  - Bouw de begeleidingsboom
    - Gebruik de ‘neighbor-joining’ methode
    - Zet de wortel op de tak waar de gemiddelde taklengte is gelijk aan beide kanten
    - Gebruik de boom om het gewicht van iedere sequentie te bepalen
  - Aligneer progressief langs de boom (seq.-seq., seq.-groep, groep-groep)
    - Vergrendel reeds gealigneerde sequenties samen
    - Aligneer twee groepen via dynamisch programmeren met sum-of-pairs scores
  - Gebruik enkele heuristieken

‘Neighbor-joining’ methode

- Afstanden d_{ij} tussen sequenties i en j worden afgeleid uit de paarsgewijs aligneringen

- Definieer de genormalizeerde afstanden D_{ij} (om te compenseren voor variabele evolutietijden) als

$$D_{ij} = d_{ij} - (r_i + r_j)$$
$$r_i = \Sigma_k \, d_{ik}/(N - 2)$$

- Initialisatie
  - De sequenties zijn de bladeren L van de boom T
- Iteratie
  - Pak het paar (i,j) waarvoor D_{ij} minimaal is
  - Definieer een nieuwe knoop k en voeg die toe aan de boom met takken

$$d_{ik} = (d_{ij} + r_j - r_j),\ d_{jk} = d_{ij} - d_{jk}$$

  - Verwijder i en j uit de bladeren L, vervang die door k en zet

$$d_{km} = (d_{im} + d_{jm} - d_{ij})/2 \qquad \text{voor alle } m \text{ in } L$$

- Eind
  - Verwijder de laatste twee bladeren en zet hun takken op d_{ij}
  - Zet de wortel op de tak waar de gemiddelde taklengte gelijk is aan beide kanten

Samenvatting

- Meervoudige alignering
- PSI-BLAST
- CLUSTALW