

Bioinformatik

Profilalignment

Profile HMMs

CLUSTAL W

Ulf Leser

Wissensmanagement in der
Bioinformatik



Definition

- Bisher
 - Vergleich zweier Strings (bzw. String und Datenbank)
- Jetzt
 - Multipler Stringvergleich: Vergleich von $k > 2$ Strings
- Definition
 - *Ein **multiple Sequenzalignment (MSA)** von k Strings S_i , $1 \leq i \leq k$, ist eine Tabelle mit k Zeilen und l Spalten, so dass*
 - *In Zeile i steht String S_i , mit beliebig eingefügten Leerzeichen*
 - *Jedes Zeichen jedes S_i steht in exakt einer Spalte*
 - *In keiner Spalte stehen nur Leerzeichen*
- Bemerkungen
 - Direkte Generalisierung des Alignment zweier Strings
 - Es folgt, dass $l = |\text{MSA}| \leq \sum(|S_i|)$
 - Warum?

Motivation II

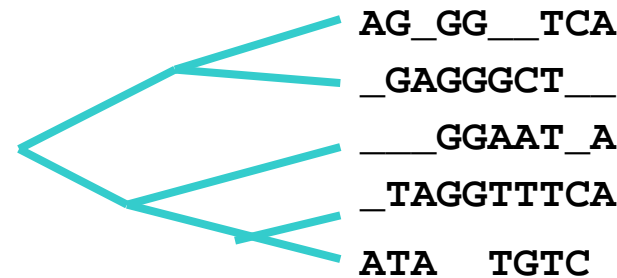
- Proteine (und damit auch DNA) setzen sich aus **funktionalen Blöcken** und „Zwischenraum“ zusammen
 - Die Blöcke findet man nicht, wenn man Sequenzen nur paarweise vergleicht
 - Bzw. man kann sie nicht vom Rauschen unterscheiden
- Trennung des eventuell zufällig Gemeinsamen (Alignment) vom bedeutungsvoll Gemeinsamen (MSA)

```
AAC  GTG  AT  T  GAC  _  
_TCGAGTGC  TTTACA  GT
```

```
AAC  GTG  AT  T  GAC  _  
_TCGAGTGC  TTTACA  GT  
GCCG  TGC  TA  GTCG  _  
TTC  AGTGGACGTG  GTA  
G  GTGCA  TGACC  _
```

MSA Zielfunktion

- **Zielfunktion** beim einfachen Alignment war klar
 - Möglichst wenig I,R,D
 - Eventuell mit Substitutionsmatrix
 - Eventuell mit spezieller Behandlung von Gaps
- Zielfunktion für MSA ist nicht so klar
 - Score einer Spalte mit 2 T, zwei G und einem Leerzeichen?
 - Angabe einer Substitutionsmatrix für k Sequenzen über Alphabet Σ würde $O(|\Sigma|^{k+1})$ Werte erfordern
 - Nicht machbar und biologisch nicht begründbar



Definitionen

- Definition

- Gegeben ein MSA M für Sequenzen S_1, \dots, S_k . Das *durch M induziertes Alignment für zwei Sequenzen S_i und S_j* ist das folgende:
 - Entferne aus M alle Zeilen außer i und j
 - Entferne alle Spalten, die in i und j ein Leerzeichen enthalten
- Gegeben ein MSA M für Sequenzen S_1, \dots, S_k . Der *Sum-Of-Pairs Score für M (SP-Score)* ist die Summe aller Alignmentsscores der durch M induzierten paarweisen Alignments
- Das *SP-Alignment Problem für Sequenzen S_1, \dots, S_k* sucht das MSA M mit minimalem SP-Score

- Bemerkung

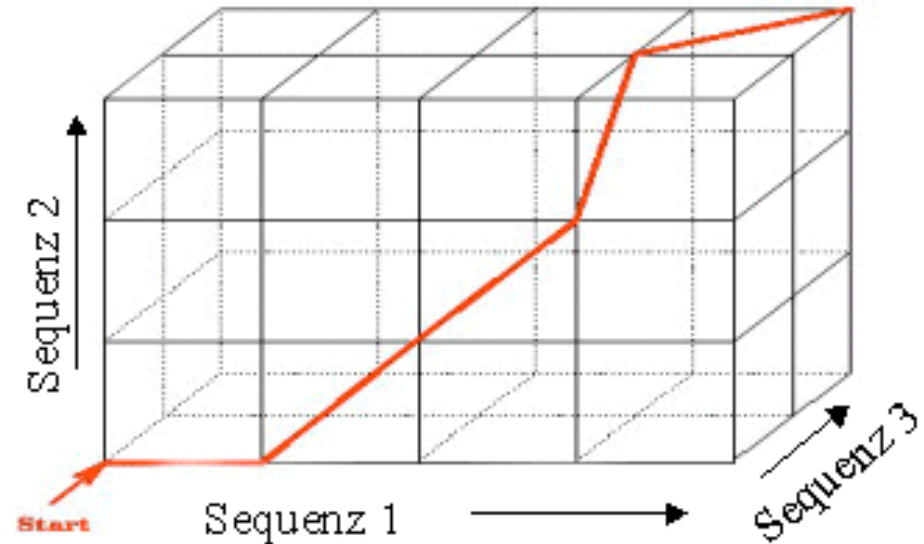
- Vergleich aller Sequenzen mit allen anderen Sequenzen – aber entsprechend dem vorgegebenen MSA

Dynamische Programmierung in k Dimensionen

- $k = 2$
 - 2-dimensionale Matrix

	0	1	2	3	4	5	6	7
		w	r	i	t	e	r	s
0	0	1	2	3	4	5	6	7
1	v	1	1	2	3	4	5	6
2	i	2	2	2	2	3	4	5
3	n	3	3	3	3	3	4	5
4	t	4	4	4	4	3	4	5
5	n	5	5	5	5	4	4	5
6	e	6	6	6	6	5	4	5
7	r	7	7	6	7	6	5	4

- $k = 3$
 - 3-dimensionale Matrix

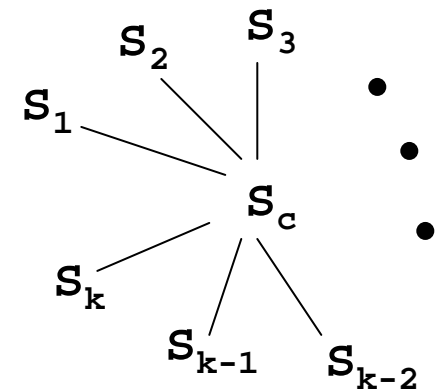


Komplexität

- Für drei Sequenzen der Länge n
 - Würfel hat n^3 Zellen
 - Für jede Zelle sind 7 Berechnungen notwendig
 - Zusammen $O(7 * n^3)$
- Allgemeiner Fall: k Sequenzen der Länge n
 - Hyperwürfel hat n^k Zellen
 - Für jede Zelle sind $2^k - 1$ Berechnungen notwendig
 - Alle Ecken eines k -dimensionalen Würfels minus eins
(Das ist die Ecke die gerade berechnet wird)
 - Zusammen $O(2^k * n^k)$
- Tatsächlich gilt
 - *Das SP-Alignment Problem ist NP vollständig*

MSA mit Konsensussequenz

- SP minimiert die Summe aller paarweisen Alignments
- Alternativ: Minimierung der **Summe der Alignments aller Sequenzen S_1, \dots, S_k mit einer Konsensussequenz S_c**
 - S_c kann eine der S_i sein, muss aber nicht
 - Konstruktion von S_c z.B. durch Untereinanderschreiben der S_i ohne Gaps und Wahl des häufigsten Buchstaben
 - MSA wird aus dem „Star“ abgeleitet
- Bei geeigneter Wahl von S_c und dem MSA – Ableitungsverfahren gilt:
 - Der SP Score des berechneten MSA ist **höchstens doppelt so hoch wie der SP-optimale**
 - Center-Star-Score approximiert den SP-Score bis auf Faktor 2



Center-Star Verfahren

- Gegeben k Sequenzen der Länge n
- Wähle als Konsensus die Sequenz S_i , die den **kleinsten durchschnittlichen Abstand** zu allen Sequenzen hat
 - Wie komplex ist das?
- Benutze S_i als Kern eines MSA M
 - Bzw. als Zentrum des Stars
- Iteriere
 - Wähle eine noch nicht alignierte Sequenz T
 - **Aligniere M und T**
 - Mit einer einfachen Methode, die wir nicht ausführen (Gusfield, p. 348)
 - Bis alle Sequenzen in M enthalten sind
- Beispiel für ein **progressives MSA** Verfahren
 - Sukzessive Sequenzen zu einem wachsenden MSA hinzufügen

Inhalt dieser Vorlesung

- Suche mit einem MSA
 - Profilalignment
 - Profile-HMM
- CLUSTAL W: Heuristisches, progressives Alignment

Suche mit MSA

- Erinnerung: Erzeugung von Proteinfamilien
 - Starte mit Proteinen gleicher/ähnlicher Funktion
 - Finde das Gemeinsame durch MSA
 - Suche „damit“ nach weiteren Vertretern
 - Modifiziere Familie entsprechend
 - Iteriere, bis Zufriedenheit eintritt
- Wie sucht man mit einem MSA?
 - Wir müssen entscheiden, wie gut eine (neue) Sequenz S zu einem MSA M passt

Variante 1: Reguläre Ausdrücke

A C A - - - A T G

T C A A C T A T C

A C A C - - A G C

A G A - - - A T C

A C C G - - A T C

Beispiel aus [Kro98]

Krogh, A. (1998). An Introduction to HMM for Biological Sequences.

$[AT] [CG] [AC] [ACGT]^* A [TG] [GC]$

- Vorteil: Schnell berechnet, schnelle Matchingalgorithmen
- Nachteil: Keine Unterscheidung der **Güte eines Matches**

Keine Unterscheidung

A C A - - - A T G

T C A A C T A T C

A C A C - - A G C

A G A - - - A T C

A C C G - - A T C

[AT] [CG] [AC] [ACGT]* A [TG] [GC]

Zwei Matches: T G C T - - A G G **Eigentlich schlecht**
 A C A C - - A T C **Bestmöglich**

Variante 2: Profile

- Definition

Gegeben ein MSA M mit n Spalten, $\Sigma' = \Sigma \cup \{-\}$

- *Das **Profil P** zu M ist eine Tabelle der Größe $n * |\Sigma'|$*
- *In der Zelle (i, j) steht die **relative Häufigkeit des Zeichens j in der Spalte i***

- Beispiel (Alphabet A, G, C)

S ₁	A	G	C	-	A
S ₂	A	G	A	G	A
S ₃	A	C	C	G	-
S ₄	C	G	-	G	C
A	0.75	0	0.25	0	0.50
G	0	0.75	0	0.75	0
C	0.25	0.25	0.50	0	0.25
-	0	0	0.25	0.25	0.25

Profile und Sequenzen

- Mit Hilfe von Profilen wollen wir Sequenzen in Bezug auf das MSA detailliert bewerten
 - Ähnlichkeit einer Sequenz zum MSA
- Das setzt wieder ein Alignment voraus
 - Welche Zeichen der Sequenz sollen mit welchen Spalten des MSA verglichen werden?
- Gegeben Sequenz S und Profil P. Wir brauchen zwei Dinge
 - Methode zur Bewertung eines konkreten Alignments von S und P
 - Methode zum Finden des besten Alignments zwischen S und P gegeben die Bewertungsfunktion

Bewertung eines Alignments

- Definition

Gegeben ein Profil P mit n Spalten, eine Sequenz S und eine Substitutionsmatrix m .

- Ein *Alignment* A von P und S ist ein Untereinanderschreiben von P und S , wobei immer eine Spalte von P (oder ein Leerzeichen) über einem Zeichen von S (oder einem Leerzeichen) steht.
 - Aber niemals zwei Leerzeichen untereinander stehen
- Wir erzeugen aus P ein P' , in dem wir an den betreffenden Stellen leere Spalten einfügen; dito ein S' aus S
- Der *Score* $s(A)$ von A berechnet sich als

$$s(A) = \sum_{i=1}^{|A|} \sum_{c_k \in \Sigma'} \begin{cases} (P'[c_k, i] * m[c_k, S'[i]]), & \text{wenn } i \text{ keine Leerspalte} \\ m[_, S'[i]], & \text{sonst} \end{cases}$$

Beispiel

Profil P

	1	2	3	4	5
A	0.75		0.25		0.50
G		0.75		0.75	
C	0.25	0.25	0.50		0.25
-			0.25	0.25	0.25

Substitutionsmatrix m

	A	G	C	-
A	2	-1	-3	-1
G		2	-1	-1
C			2	-1

Sequenz S: AAGGC

Alignment A

1	-	2	3	4	5
A	A	G	-	G	C

$$\begin{aligned}
 s(A) &= (2*0.75 + -1*0 + -3*0.25 + -1*0) + \\
 &\quad (-1) + \\
 &\quad (-1*0 + 2*0.75 + -1*0.25 + -1*0) + \\
 &\quad (-1*0.25 + -1*0 + -1*0.50 + -1*0.25) + \\
 &\quad (-1*0 + 2*0.75 + -1*0 + -1*.25) + \\
 &\quad (-3*0.50 + -2*0 + 2*0.25 + -1*0.25) \\
 &= 0.75 - 1 + 1.25 - 1 + 1.25 - 1.25 \\
 &= 0
 \end{aligned}$$

Optimale Profilalignments

- Wie finde ich das optimale Profilalignment?
 - Natürlich mit **dynamischer Programmierung**

- Theorem

Gegeben eine Substitutionsmatrix m , Profil P , Sequenz S

- *Sei $c(x, j)$ der Score für das Alignieren eines Zeichen x mit Spalte j in P , also*

$$c(x, j) = \sum_{c_k \in \Sigma'} P[c_k, j] * m[c_k, x]$$

- *Sei $v(i, j)$ der Score für das optimale Alignment von den ersten i Spalten von P mit dem Präfix $S[1..j]$. v berechnet sich als*

$$v(i, j) = \max \left(\begin{array}{l} v(i-1, j) + c(_, i) \\ v(i, j-1) + m[_, j] \\ v(i-1, j-1) + c(S[j], i) \end{array} \right)$$

PSI-BLAST

- Implementierung der iterativen Suchstrategie
 - Gegeben Suchsequenz S: Berechne Profil P
 - Durchsuche DB mit P
 - Tatsächlich wird eine lineare Variante ohne Gaps verwendet
 - Bilde ein multiples Alignment von allen Hits
 - Berechne daraus eine neues P
 - Iteriere, bis Stoppkriterien erfüllt
- **Erhöhung der Sensitivität** gegenüber einfachem BLAST
 - PSI-BLAST findet auch weiter entfernte Homologien
 - Zwitter zwischen Patternmatching / Homologiesuche

PSI-BLAST und Duplikate

- Problem: **Redundante Sequenzen**
 - n-fach vorkommende Sequenzen, z.B. ESTs
 - Gaukeln **falsche Positionsgewichte** vor
 - Base ist nicht evolutionär häufig, sondern wurde nur sehr oft gemessen worden
- Lösung: Entfernung fast-identischer Sequenzen
 - Reine Heuristik
 - Auf Sequenzebene kann man nicht unterscheiden
 - Redundante Sequenzen (identisches EST zweimal sequenziert, womöglich mit unterschiedlichen Sequenzierfehlern)
 - Nicht redundante, sehr ähnliche Sequenzen (homologes Gen bei Maus und Mensch, Genduplikate)

Inhalt dieser Vorlesung

- Suche mit einem MSA
 - Profilalignment
 - Profil-HMM
- CLUSTAL W: Heuristisches, progressives Alignment

Profil-HMM

- Wir verfolgen einen ähnlichen Gedanken wie beim Schritt vom **Alignment zum Paar-HMM**
 - Paar-HMM: Eher selten genutzt, da keine schnellen Suchalgorithmen etabliert und schlechte Hits oftmals nicht interessieren
 - Profil-HMM: Etablierte und häufig verwendete Methode
 - Berechnung der **Gesamt-Wsk** einer Sequenz S gegeben ein MSA ist wesentlich **sensitivere Methode** als nur der beste Pfad
 - Beste Pfade sind oftmals relativ schlecht (da „echte“ Matches zwischen MSA-Spalte und Zeichen von S sehr selten)
 - Oft gibt es viele ungefähr gleich gute Alignments

Erster Schritt: Ein spezielles HMM

- Idee
 - Jede „hinreichend volle“ Spalte wird ein **Match-Zustand** des HMM
 - „Hinreichend voll“: Wählbarer Schwellwert für die Anzahl an INSELS
 - Seine Emissionswahrscheinlichkeiten entsprechen den Häufigkeiten der Zeichen in der Spalte
 - „Halbvolle, nicht vollkommen leere“ Spalten darf man (muss man aber nicht) überspringen: **Insertions**
 - „Fast-leere“ Spalten lassen wir gleich ganz weg
 - Die Übergangs-Wsk entnehmen wir dem MSA

MSA – HMM

```

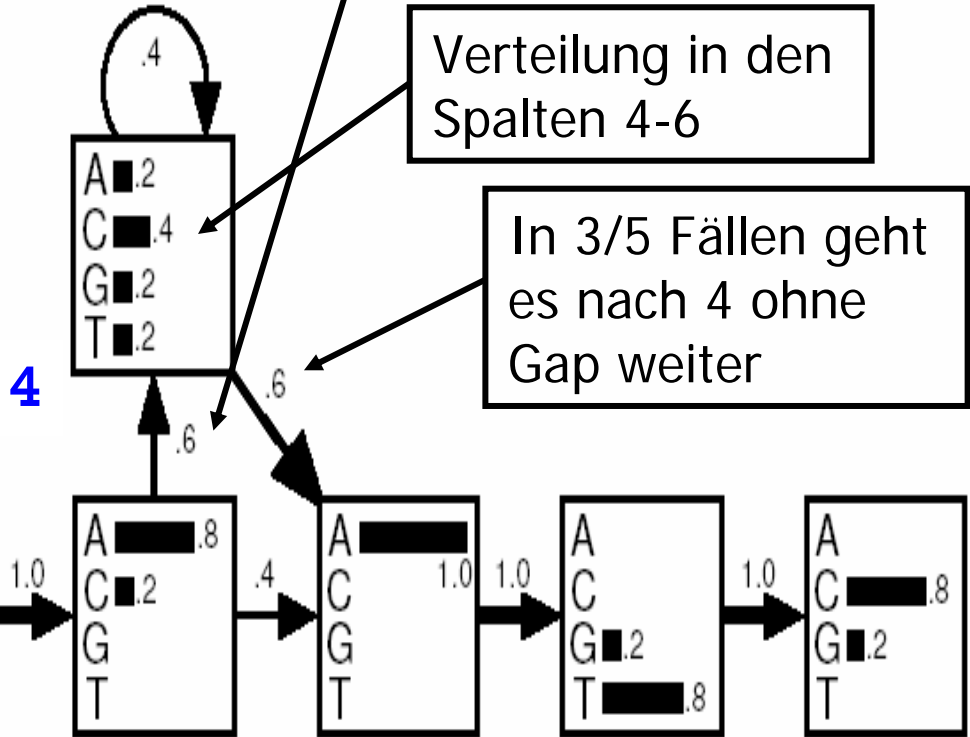
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
1 2 3 4 5 6 7 8 9
    
```

Nach Spalte 3 betritt man in 3/5 Fällen den INS State (und macht nicht gleich mit 7 weiter)

Insertion

Verteilung in den Spalten 4-6

In 3/5 Fällen geht es nach 4 ohne Gap weiter

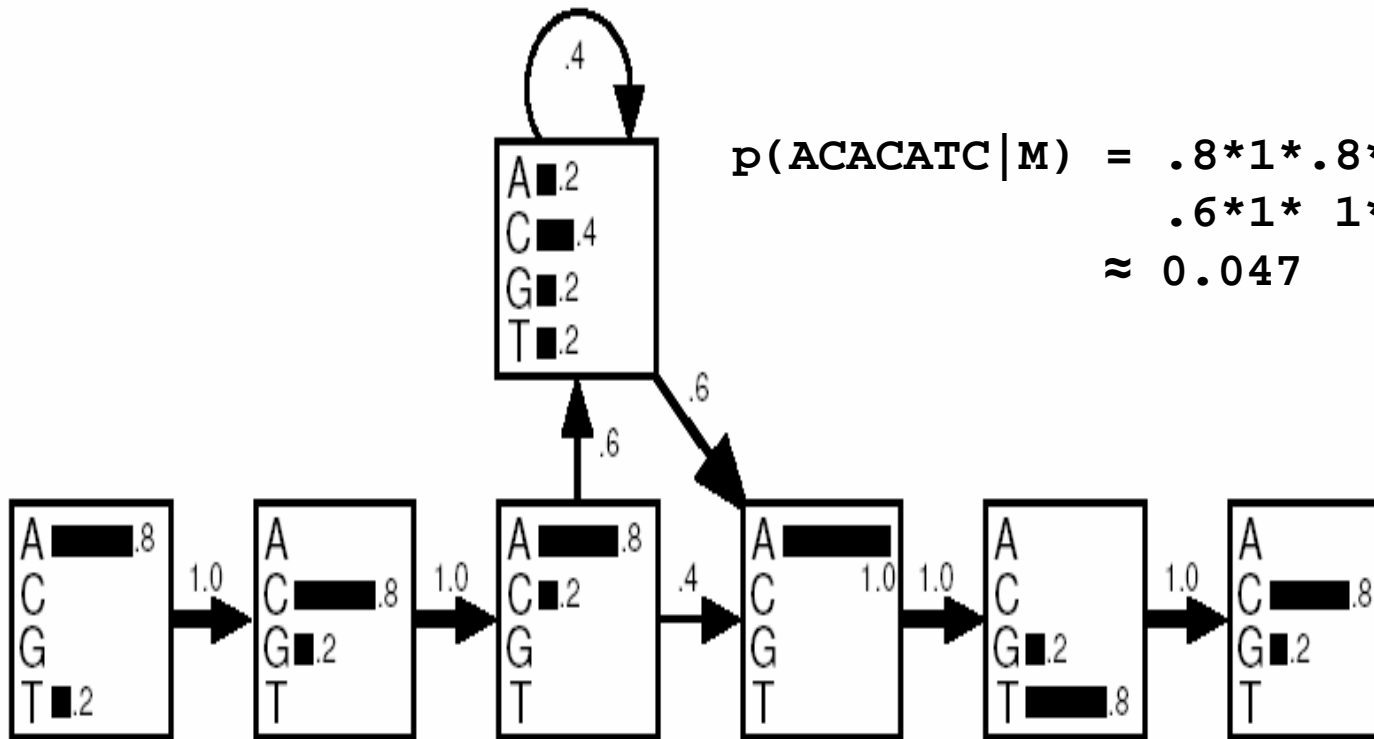


Matches

[Kro98]

Scoring

- Wir können nun die maximale Wsk einer neuen Sequenz, gegeben das Modell, berechnen
 - Sprich: Zu der Sequenz-/ Proteinfamilie zu gehören



$$p(\text{ACACATC} | M) = .8 * 1 * .8 * 1 * .8 * .6 * .4 * .6 * 1 * 1 * .8 * 1 * .8 \approx 0.047$$

Beispielscores (Bester Pfad)

	Sequenz	Wsk (%)
Consensus	ACAC--ATC	4.7
Beispiel 1	ACA---ATG	3.3
Beispiel 2	TCAACTATC	0.0075
Beispiel 3	ACAC--AGC	1.2
Beispiel 4	AGA---ATC	3.3
Beispiel 5	ACCG--ATC	0.59
Eigentlich schlecht	TGCT--AGG	0.0023

- Problem
 - Score hängt sehr stark von der Länge der Sequenz ab
- Lösungen
 - Normalisieren
 - Log-Odds Score wählen

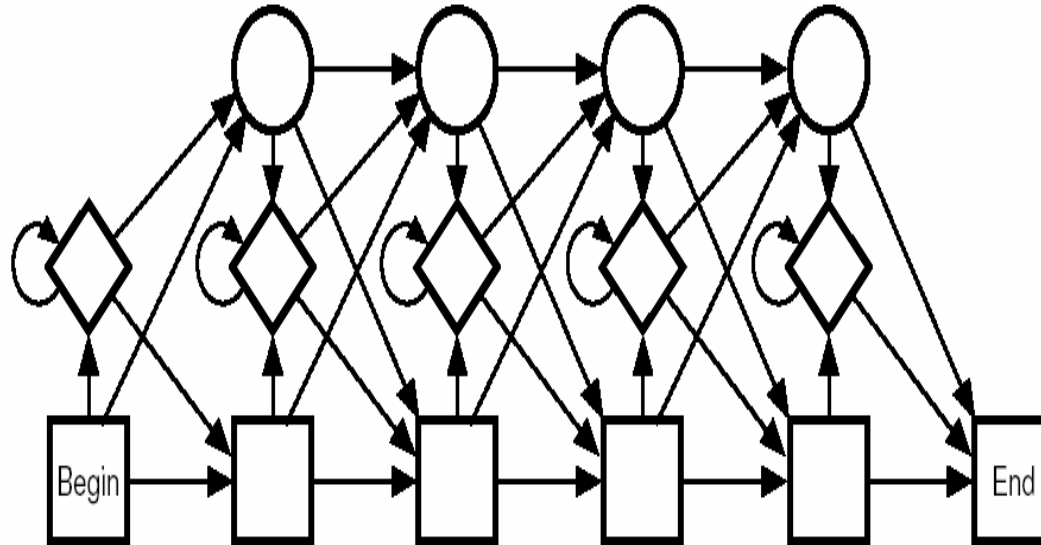
Log Odds Scores

	Sequenz	Wsk (%)	Log-odds
Consensus	ACAC--ATC	4.7	6.7
Beispiel 1	ACA---ATG	3.3	4.9
Beispiel 2	TCAACTATC	0.0075	3.0
Beispiel 3	ACAC--AGC	1.2	5.3
Beispiel 4	AGA---ATC	3.3	4.9
Beispiel 5	ACCG--ATC	0.59	4.6
Eigentlich schlecht	TGCT--AGG	0.0023	-0.97

Profil-HMMs

- Das eben angewandte Verfahren zur Umwandlung einer MSA in ein HMM war doch sehr heuristisch
 - Warum keine Insertions an anderen Stellen erlauben? Die sollten zwar bestraft werden, aber möglich sein
 - Warum keine Deletions? Müssen auch möglich sein
- Profil-HMMs: **HMM mit spezieller Struktur**
 - Feste, repetitive Zustands-Struktur
 - (Un-)Wsk der INSDELS wird durch die **Übergangs-Wsk** ausgedrückt

Struktur

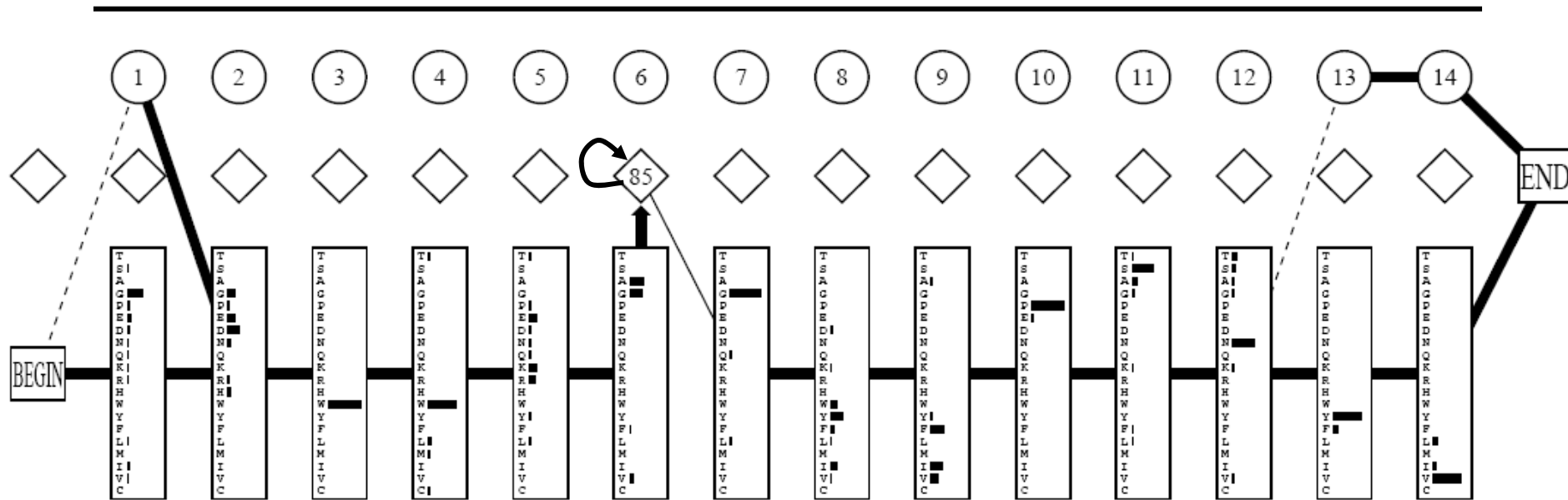


- Rechtecke: Match-Zustände
 - Symbolisieren relativ volle Spalten (Typisch: >50%)
- Rauten: Insertion-Zustände
 - Symbolisieren **Spalten/Bereiche mit vielen Gaps**
 - Können beliebig oft wiederholt werden
- Kreise: Deletion-Zustände
 - Überspringen jeweils genau einen Match-Zustand, aber können verkettet sein
 - „**Silent States**“: Emittieren keine Zeichen

Beispiel

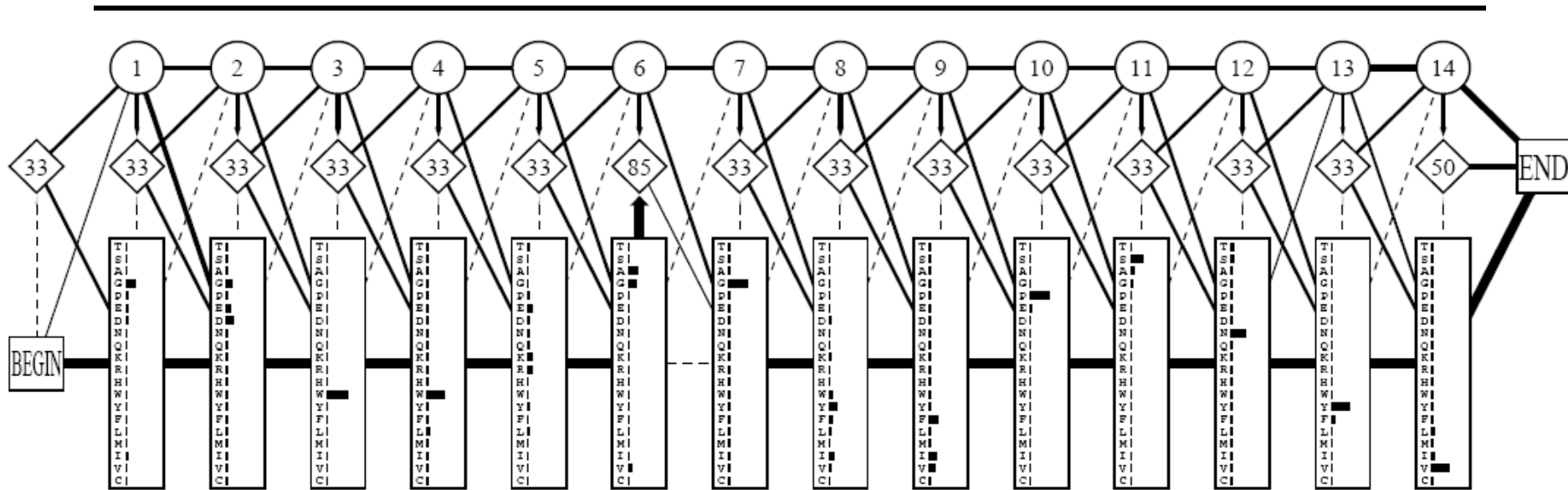
```
GGWWRRGdy.ggkkkqLWFPSSNYV
IGWLNNGdyne.ttgkerLGDFPSTYV
PNWWEQGql..nnrrrGIFPSTNYV
DEWWEQAarr..deqqiGIVPSK--
GEWWEKAqrs..tggqqeGFI PFNFV
GDWWLArss..sggqqtGYIPSTNYV
GDWWDAel..kqgrrrGKVPSNYL
-DWWEArsslssghrGYVPSNYV
GDWWYArsslitnseGYIPSTYV
GEWWEKArsllatrkGYIPSTNYV
GDWWLArslsvtgreGYVPSNFV
GEWWEKAkslsskreGFI PSNYV
GEWCEAqt.knngq.GWVPSNYI
SDWWRVvnltttrqqeGLIPLNFV
LPWWRARrd.knngqqeGYIPSTNYI
RDWWEFRsktvtytpGYYESGYV
EHWWEKVkd.algnvGYIPSTNYV
IHWWRVqrd.rnqgheGYVPSNYL
KDWWEKVe.v..ndrqqGFPAAAYV
VGWMPGlnerttrqqrGGDFPSTYV
PDWWEGgel..nggqqrGVFPASV
ENWWEWNGeie..gnrkGIFPATYV
EEWLEGEcec..kqgkvGIFPKVFV
GGWWEKGDy.gttriqqQYFPSNYV
DGWWRGsy..nggqvGWFPSTNYV
QGWWRGgei..ygrvGWFPANV
GRWWEKAarr..angetGIIPSTNYV
GGWWEAarr..tngenGYIPSTNYV
NDWWTGr.t..nqkeGIFPANV
```

Das Profil-HMM dazu



- Schlecht konservierter Block wird ein einziger INS Zustand
 - „85“: Wsk des Übergangs zum selben Zustand (Kreis)
 - Schätzen der Parameter: Später
- Offensichtliches Problem: **Overfitting**
 - Deletions sind praktisch überall „verboten“

Profil-HMM mit Pseudo-Counts



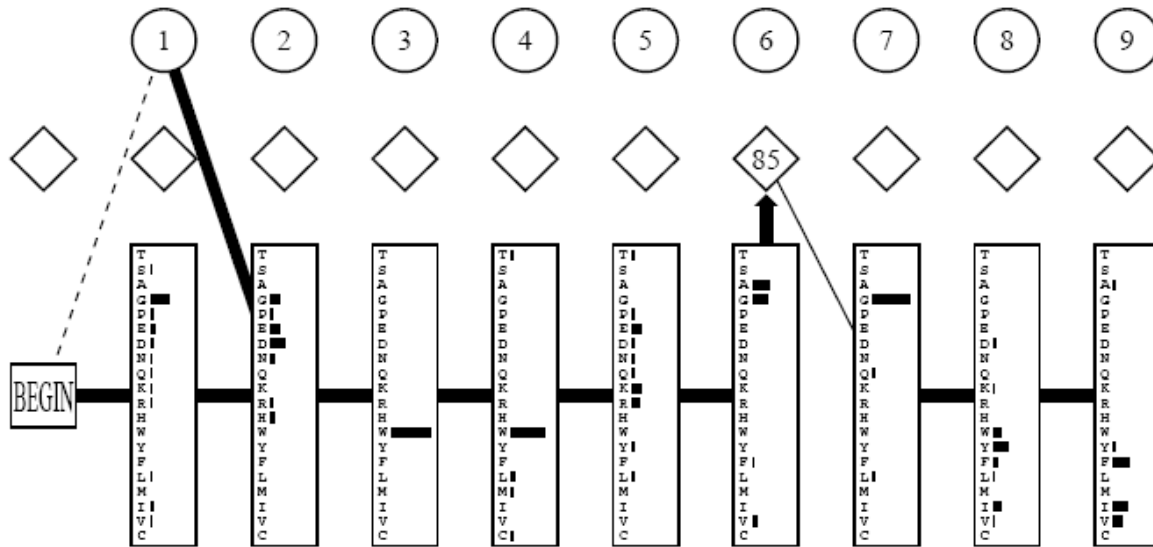
- Profil-HMM, gelernt mit einem Pseudo-Count von 1
- **Smoothing** ist sehr wichtig, da Sequenzfamilien selten 1000de Sequenzen umfassen
 - Und as Profil-HMM sehr viele Zustände hat

Lernen eines Profile-HMM

- Wir haben den (glücklichen) Fall 1 des Trainingsproblems für HMM
 - Alles Sequenzen des MSA sind im MSA aligniert
 - Damit **wissen wir zu jeder Sequenz genau die Zustandsfolge**
 - Wir können also eine einfache Maximum Likelihood Schätzung vornehmen (hier nur ohne Pseudo-Counts)

$$a_{st} = p(t | s) = \frac{A_{st}}{\sum_{t' \in M} A_{st'}} \quad e_s(x) = \frac{E_s(x)}{\sum_{x' \in \Sigma} E_s(x')}$$

Profile-HMM erklärt



G	G	W	W	R	G	d	y	.	g	g	k	k	q	L	W	F	P	S	N	Y	V
I	G	W	L	N	G	d	y	.	g	g	k	k	q	L	W	F	P	S	N	Y	V
P	N	W	W	E	G	q	r	.	n	n	r	r	i	G	I	F	P	S	N	Y	V
D	E	W	W	Q	A	r	r	.	d	e	q	q	i	G	I	F	P	S	N	Y	V
G	D	W	W	K	L	a	r	.	s	g	g	q	e	G	Y	I	P	S	N	Y	V
G	D	W	W	D	A	r	e	.	s	k	g	g	r	G	K	V	P	S	N	Y	L
-	D	W	W	E	A	r	s	.	l	s	g	g	h	G	Y	V	P	S	N	Y	V
G	D	W	W	Y	A	r	s	.	l	s	t	n	s	G	Y	I	P	S	T	Y	V
G	G	D	W	W	K	A	r	.	s	l	a	t	r	G	Y	I	P	S	N	Y	V
G	G	D	W	W	L	A	k	.	s	l	v	s	t	G	Y	I	P	S	N	F	V
G	G	E	W	W	C	E	A	.	q	t	.	k	n	G	W	V	P	S	N	Y	I
S	D	W	W	R	V	v	n	.	l	t	t	r	g	G	L	I	P	L	N	F	V
L	P	W	W	R	A	r	r	.	k	t	n	g	g	G	Y	I	P	S	N	Y	I
E	H	W	W	K	V	k	d	.	.	a	l	g	g	G	Y	I	P	S	S	Y	L
I	K	D	W	W	K	V	q	.	e	.	r	n	d	G	Y	I	P	S	S	Y	L
V	G	W	W	M	P	V	g	.	e	.	.	n	r	G	D	F	P	G	T	Y	V
P	D	W	W	E	G	g	e	.	.	.	n	g	g	G	V	F	P	A	S	Y	V
E	E	W	W	L	E	G	e	.	.	.	g	g	n	G	I	F	P	K	V	F	V
G	G	W	W	K	G	d	y	.	g	.	t	r	r	G	I	F	P	S	N	Y	V
D	G	W	W	R	G	s	e	.	.	.	n	g	g	G	W	F	P	P	A	N	Y
G	G	R	W	W	K	A	r	.	.	.	a	n	s	G	I	I	P	P	S	N	Y
G	G	D	W	W	T	Q	G	.	.	.	k	s	t	G	W	A	P	T	N	Y	L
N	D	W	W	T	G	r	t	.	.	.	n	.	.	G	I	F	P	A	N	Y	V

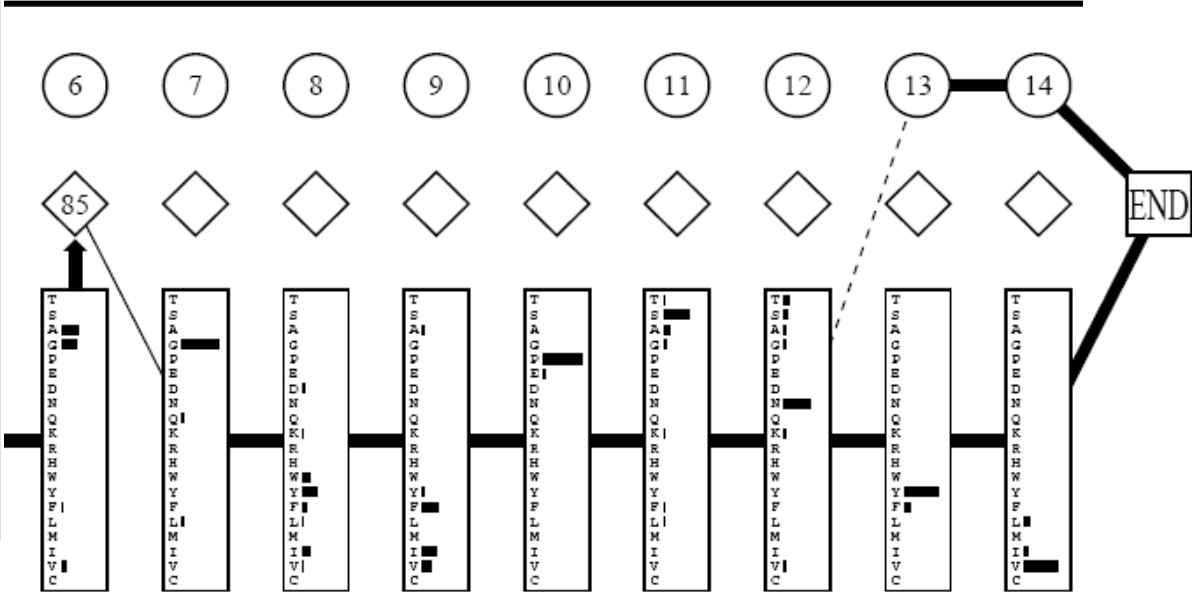
- Erste Spalte: 1 Gap, also Wsk 1/30 für Start-DEL und 29/30 für Start-Match1
- Keine Gaps in Spalten 2-6
- Nach Match6 kommt mit 100% Wsk eine Insertion, keine Sequenz hat nur Leerzeichen zwischen den beiden „guten“ Blöcken
- Zählen aller Übergänge INS-INS und INS-Match7 bringt das Verhältnis 85/100

Profile-HMM erklärt

```

GGWLNRRGGynee.ttgkkeeqrLWFFPPSSNTYVV
GPNWVWGAqrll...nngkrrGGIFPPSSNYV
DEWVWQAARrrs...dqqkrrGGIVPPSSK-V
GGDWWLAArsl...stqqqrGGYIIPPSSNYVV
GGDWWDAARsl...skqqqrGGKVPSSNYVL
-DWWEAARsl...stqqqrGGYVPPSSNYVV
GGDWWVAAARsl...stqqqrGGYIIPPSSNYVV
GGDWWLAAARsl...stqqqrGGYVPPSSNYVV
GGEWVWKAARsl...stqqqrGGFVPPSSNYVV
GGEWVWKAARsl...stqqqrGGYIIPPSSNYVV
SSDPVWRAVnd...ltnnqqqrGGGLIPLNFV
LRDWWERFRkd...ktnvyygqqrGGYIIPPSSNYVV
IHKVWVWPKVg...lne...nrrqqqrGGDFPAGATYVV
PDDWVWEGG...l...nngqqqrGGVFPATYVV
EENWVWNGG...l...gkqqqrGGIFPKVYVV
DDGGVWVWKG...l...yqqqrGGWFPSSNYVV
OGRVWVWKAARrr...akstqqqrGGWAPSSNYVL
GGDWWVWTR...l...nngqqqrGGIIPPSSNYVV

```



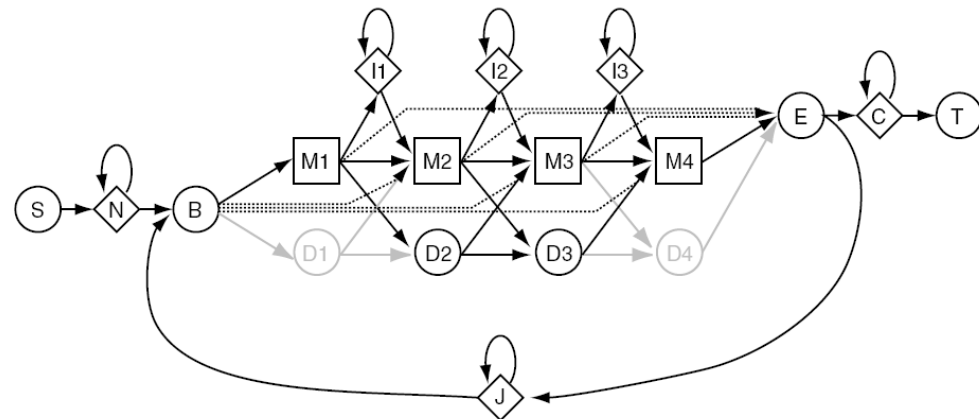
- Spalte 10 ist hoch konserviert: $p(P)=29/30$, $p(E)=1/30$
- Spalte 11 ist hoch konserviert: $p(S)=19/39$, $p(A)=5/30$, ...
- In Sequenz 4 sind am Ende zwei DELS
- Etc.

Scoring mit einem Profile-HMM

- Wie gehabt
 - Viterbi: Insgesamt wahrscheinlichstes Alignment
 - Forward/Backward: Gesamtwahrscheinlichkeit der Sequenz gegeben das Modell (das MSA)

Erweiterungen

- **Gewichtung der Einzelsequenzen** vor Berechnung der Emissions-Wsk
 - Duplikate filtern, Verteilungs-Bias bezogen auf den Stammbaum minimieren
- Profile-HMMs mit **eingebetteten** oder wiederholten MSA-Matches
 - Findet einen oder mehrere Matches für ein MSA in einer Sequenz („lokales Profilmatching“)
 - Beispiel: PLAN7 Struktur in HMMER (Eddy 98)
 - N,C,J: Random sequence emitting states



Inhalt dieser Vorlesung

- Suche mit einem MSA
 - Profilalignment
 - Profile-HMM
- CLUSTAL W: Heuristisches, progressives Alignment

CLUSTAL W

- Greedy-Variante des Alignments mit einem phylogenetischen Baum
- Lange Zeit das **Standardprogramm** zum multiplen Sequenzalignment
 - Higgins, D. G. and Sharp, P. M. (1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer." *Gene* **73**(1): 237-44.
 - Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Res* **22**(22): 4673-80.
- Heute ist die Situation unübersichtlicher
 - DAlign, T-Coffee, HMMT, PRRT, MULTALIGN, ...

Progressives Alignment

- Grundproblem des Sum-Of-Pair Scores für P-MSA
 - Ständige Betrachtung aller Sequenzen
 - Das sind zu viele Möglichkeiten = Dimensionen
- Grundidee der progressiven Verfahren
 - Berechne zunächst **MSA für Teilmengen** von Sequenzen
 - Baue das **Gesamt-MSA aus den Teil-MSA**
- Das wirft Fragen auf
 - Wie wähle ich die Teilmengen?
 - Wie verschmilzt man Teillösungen (MSA)
 - In welcher Reihenfolge verschmilzt man die Teil-MSA?
 - Erreicht man optimale Lösung?
 - Nicht die SP-optimale Lösung
 - Aber in der Praxis sehr gute Ergebnisse

CLUSTAL W: Grundaufbau

- Gegeben k Sequenzen
- Drei Schritte
 - **Ähnlichkeitsmatrix**: Berechne alle paarweisen Alignmentsscores
 - Konstruiere einen „**Guide Tree**“ durch hierarchisches Clustering
 - Berechne und verschmelze Teil-MSA gemäß dem Guide Tree
- Idee dahinter
 - Aligniere erst sehr ähnliche Sequenzen – **Signale werden verstärkt**
 - Werden z.B. zwei sehr verschiedene Cluster von Sequenzen betrachtet, berechnet CLUSTAL automatisch erst zwei (homogene) MSA und verschmilzt diese am Ende
 - Hohe Chance, dass **konservierte Blöcke** erhalten bleiben
 - **Außenseiter** kommen erst spät dazu und können die Blockstruktur nicht mehr stören
 - Orientierung an der „tatsächlichen“ Entstehungsgeschichte, dem **phylogenetischen Baum**

Schritt 1 und 2

- Berechnen der Ähnlichkeitsmatrix M
 - $O(k^2)$ Alignment scores
- Hierarchisches Clustering (ursprünglich)
 - Wähle Zelle (i,j) mit kleinstem Abstand aus Matrix M
 - Das ist das erste Paar
 - Erzeuge M': Lösche die Sequenzen i und j aus M und füge neue Spalte/Zeile (ij) ein
 - Für alle $k \neq ij$: $M'[ij,k] = (M[i,k] + M[j,k]) / 2$
 - Mittlerer Abstand zu i und j
 - Iteriere, bis Matrix nur noch 2x2 groß ist
- Tatsächlich benutzt CLUSTAL W heute Neighbour Joining
 - Details später: Ultrametrien und additive Bäume

Beispiel: Graphisch

A
B
C
D
E
F
G

ABCDEFG

A
B.
C..
D...
E....
F.....
G.....

(B,D) → a

A
B
C
D
E
F
G



ACEFGa

A
C.
E..
F...
G....
a.....

(E,F) → b

A
B
C
D
E
F
G

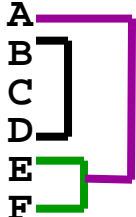


ACGab

A
C.
G..
a...
b....

(A,b) → c

A
B
C
D
E
F
G

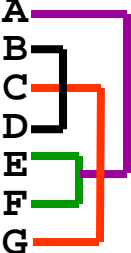


CGac

C
G.
a..
c...

(C,G) → d

A
B
C
D
E
F
G

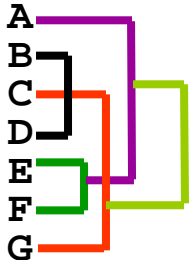


acd

a
c.
d..

(d,c) → e

A
B
C
D
E
F
G

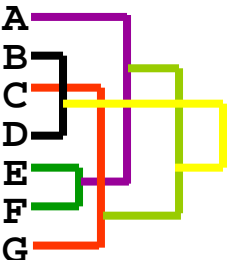


ae

a
e.

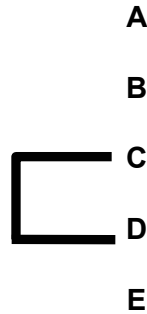
(a,e) → f

A
B
C
D
E
F
G

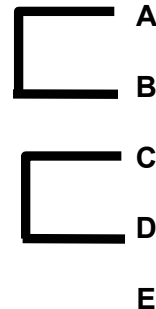


Konstruktion des Guide Trees

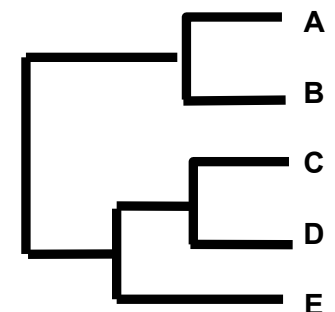
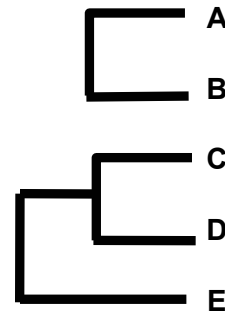
	A	B	C	D	E
A		17	59	59	77
B			37	61	53
C				13	41
D					21



	A	B	E	CD
A		17	77	59
B			53	49
E				31



	E	CD	AB
E		31	65
CD			54



Schritt 3: Progressive MSA Generierung

- Berechnung paarweiser Alignments in der Reihenfolge des **Guide Trees**
- Alignment eines MSA M_1 mit einem MSA M_2
 - Dynamische Programmierung mit linearem Gapscore
 - Wert eines Mismatches/Matches ist der **Durchschnittsscore aller Paare** mit einem Zeichen aus M_1 und einem aus M_2
 - Bei k Sequenzen müssen daher maximal $k/2 * k/2 = O(k^2)$ Scores (für einen Spaltenvergleich) betrachtet werden

- Beispiel

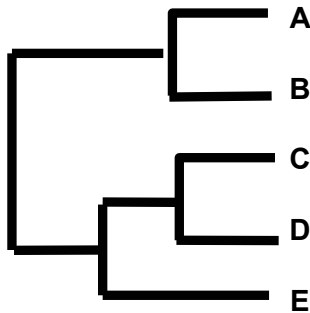
A ...P...
B ...G...
C ...P...

D ...A...
E ...A...
F ...Y...

- Score des Spaltenvergleichs

$$- (2 * m[P, A] + m[P, Y] + 2 * m[G, A] + m[G, Y] + 2 * m[P, A] + m[P, Y]) / 9$$

Beispiel



C PADKTNVKAAWGKVG**A**HAGEYGA

D AADKTNVKAAWSKVGGHAGEYGA

A PEEKSAVTALWGKVN**V**DEYGG

B GEEKAAVLALWDKVNEEEYGG

C PADKTNVKAAWG_**K**VGAHAGEYGA

D AADKTNVKAAWS_**K**VGGHAGEYGA

E AA__TNVKTAWSSKVGGHAPA__**A**

A PEEKSAV_**T**ALWG_**K**VN__VDEYGG

B GEEKAAV_**L**ALWD_**K**VN__EEYGG

C PADKTNV**K**AA_**W**G_**K**VGAHAGEYGA

D AADKTNV**K**AA_**W**S_**K**VGGHAGEYGA

E AA__TNV**K**T**A**_WSS**K**VGGHAPA__**A**

Once a gap, always a gap

Iterative Verfahren

- CLUSTAL W ist „greedy“
 - Ergebnis abhängig von der Reihenfolge des Hinzunehmens von Sequenzen
 - Der Guide Tree kann aber (evolutionär) falsch sein oder wenig Aussagekraft haben, wenn Unterschiede nur sehr klein sind
 - Was am **Anfang schief läuft, ist besonders schlimm**
 - „Once a gap, always a gap“
 - Einmal getroffene Alignments werden nie mehr korrigiert, nur „gestreckt“
- Was kann man tun?
 - Verschiedene Trees probieren und Ergebnisse vergleichen
 - Sampling – verschiedene Sequenzmengen versuchen
 - **Iterative Verfahren**
 - Sukzessive Verbesserung eines (progressiv gefundenen) Alignments
 - Jede Sequenz einmal entfernen und neu alignieren
 - Solange bis Konvergenz