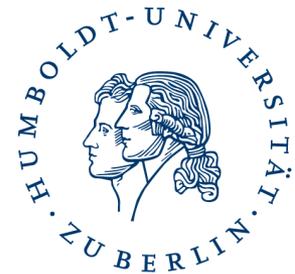


# Algorithmische Bioinformatik

CLUSTAL W  
Profilalignment  
Profile HMMs

Ulf Leser

Wissensmanagement in der  
Bioinformatik



# Inhalt dieser Vorlesung

---

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

# CLUSTAL W

---

- Greedy-Variante von „MSA mit phylogenetischem Baum“
- Lange Zeit das **Standardprogramm** für MSA
  - 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 sind viele Tools im Einsatz: DAlign, T-Coffee, HMMT, PRRT, MULTALIGN, ...

# Progressives Alignment

---

- Grundproblem des Sum-Of-Pair Scores für P-MSA
  - Ständige Betrachtung aller Sequenzen => Exponentieller Suchraum
  - Überschätzung der realen Editabstände – ein evolutionäres Ereignis wird in vielen Sequenzpaaren gezählt
- Progressive Verfahren: Berechne iterativ **MSA für wachsende Teilmengen** von Sequenzen
  - Wie wähle ich größer werdende Teilmengen?
  - Wie verschmilzt man zwei kleinere MSA zu einem größeren?
  - In welcher Reihenfolge verschmilzt man die Teil-MSA?
  - Wie gut funktioniert das?

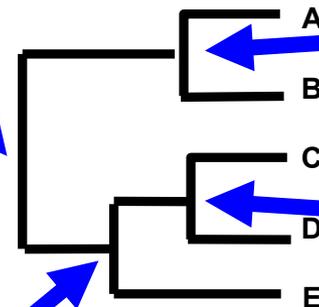
# CLUSTAL W: Grundaufbau

- Gegeben k Sequenzen. Drei Schritte
  - Ähnlichkeitsmatrix: Berechne alle paarweisen Alignmentsscores
  - Konstruiere einen „Guide Tree“
  - Berechne und verschmelze Teil-MSA gemäß dem Guide Tree

```
A PEEKSAVTALWGKVNVD EYGG
B GEEKAAVLALWDKVN EEEYGG
C PADKTNVKA AAWGKVG AHAGEYGA
D AADKTNVKA AAWSKVGG HAGEYGA
E AATNVKTA WSSKVGGHAPA A
```

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

```
A PEEKSAV TALWG KVN VDEYGG
B GEEKAAV LALWD KVN EEEYGG
C PADKTNV KAA WG KVG AHAGEYGA
D AADKTNV KAA WS KVGG HAGEYGA
E AA TNV KTA WSSKVGGHAPA A
```



```
A PEEKSAVTALWGKVNVD EYGG
B GEEKAAVLALWDKVN EEEYGG
```

```
C PADKTNVKA AAWGKVG AHAGEYGA
D AADKTNVKA AAWSKVGG HAGEYGA
```

```
C PADKTNVKA AAWG KVG AHAGEYGA
D AADKTNVKA AAWS KVGG HAGEYGA
E AA TNV KTA WSSKVGGHAPA A
```

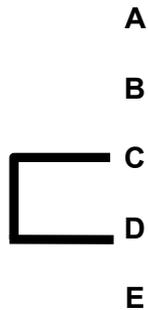
# Schritt 1 und 2

---

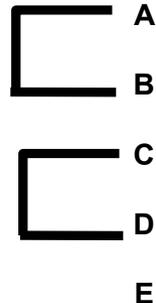
- Schritt 1: Berechnen der Ähnlichkeitsmatrix M
  - $O(k^2)$  paarweise Alignments
- Schritt 2: Hierarchisches Clustering (ursprünglich)
  - Wähle Zelle (i,j) mit kleinstem Abstand aus Matrix M
    - Das ist das erste Paar
  - Erzeuge  $M'$ : Lösche die Sequenzen i und j aus M und füge neue Spalte/Zeile (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
- Heute: Neighbour Joining (später)

# Beispiel

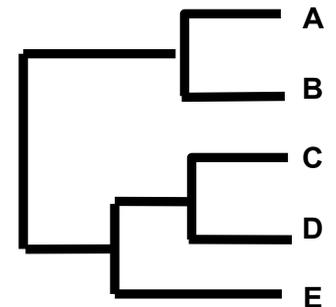
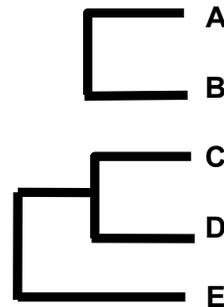
	A	B	C	D	E
A		17	59	59	77
B			37	61	53
C				13	41
D					21



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

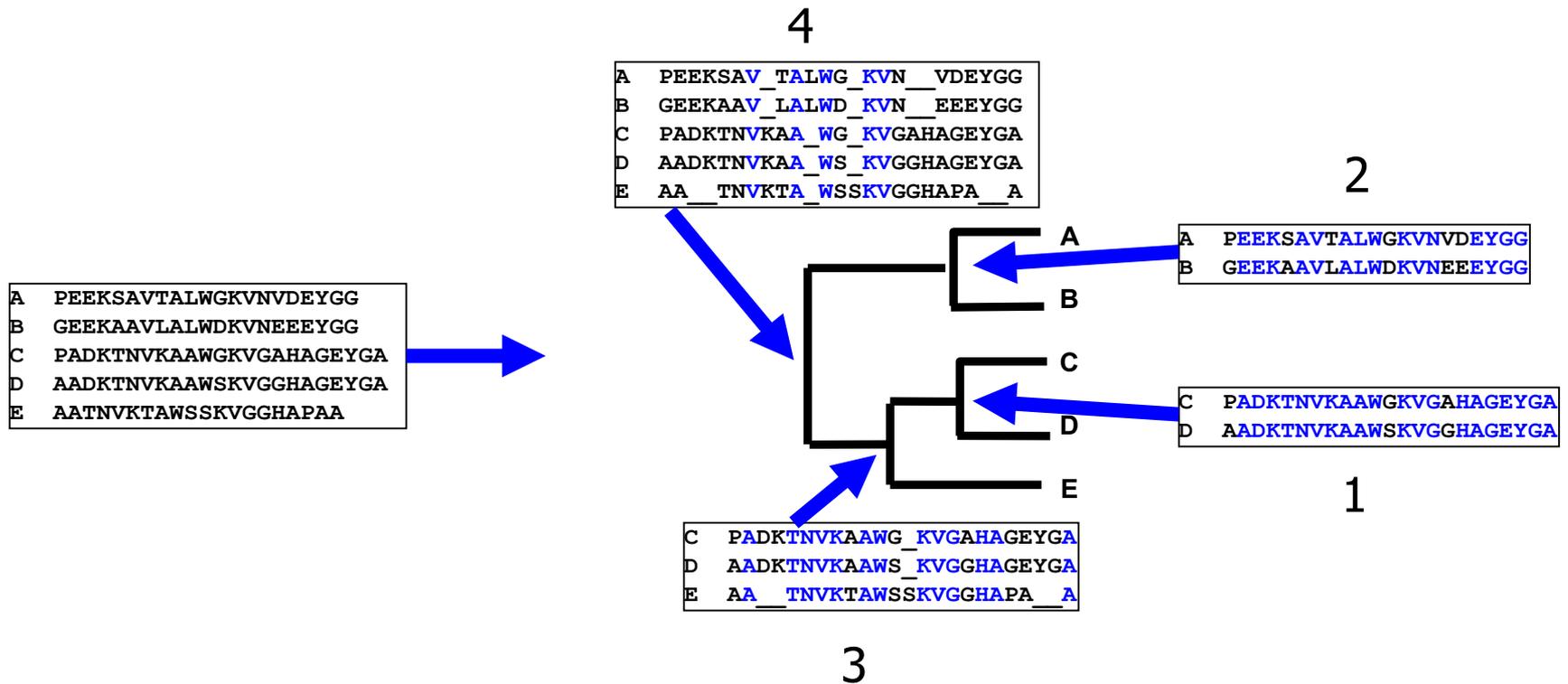


	E	CD	AB
E		31	65
CD			54



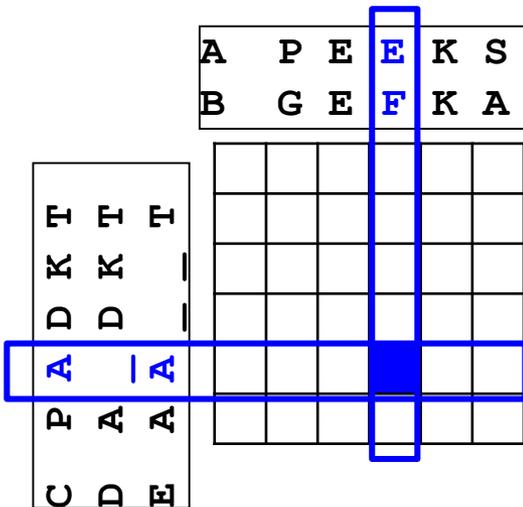
# Schritt 3: Progressive MSA Generierung

- MSA werden in Reihenfolge des Guide Trees verschmolzen



# Verschmelzen zweier MSA

- Geg. MSA  $M_1$  mit  $k_1$  Sequenzen und MSA  $M_2$  mit  $k_2$  Seq.
- Berechnung eines Alignment über den Spalten von  $M_1/M_2$ 
  - Wert eines Spaltenpaars ist der **Durchschnittsscore aller Paare** mit einem Zeichen aus der  $M_1$ -Spalte und einem aus  $M_2$ -Spalte
  - Braucht  $k_1 * k_2 \in O(k^2)$  Zeichenvergleiche (pro Spaltenpaar)



Score des Spaltenvergleichs:

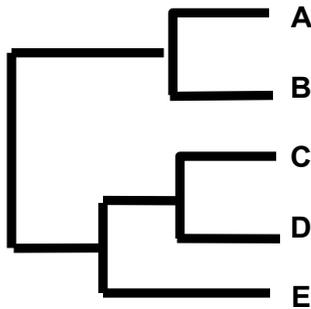
$$(2 * m[A, E] + m[_ , E] + 2 * m[A, F] + m[_ , F]) / 6$$

# Idee: Guide Tree + Progressives Alignment

---

- Aligniert erst sehr ähnliche Sequenzen – **Konservierte Bereiche werden erhalten**
  - Existieren z.B. sehr unterschiedliche Cluster, berechnet CLUSTAL automatisch erst (homogene) MSA und verschmilzt diese spät
  - Hohe Chance, dass **konservierte Blöcke** erhalten bleiben
- **Außenseiter** kommen spät dazu und zerstören die Gesamtstruktur des MSA nicht mehr
- Orientierung an der „tatsächlichen“ Entstehungsgeschichte, dem **phylogenetischen Baum**

# Beispiel



C PADKTNVKAAWGKVG**A**HAGEYGA  
D AADKTNVKAAWSKVGGHAGEYGA

A PEEKSAVTALWGKVN**V**DEYGG  
B GEEKAAVLALWDKVNEEEYGG

C PADKTNVKAAWG\_ **K**VGAHAGEYGA  
D AADKTNVKAAWS\_ **K**VGGHAGEYGA  
E AA\_ **T**NVKTAWSSKVGGHAPA\_ **A**

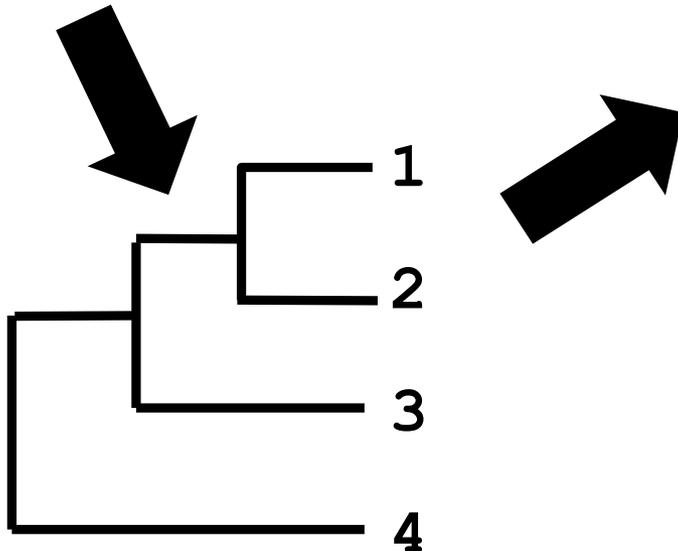
A PEEKSAV\_ **T**ALWG\_ **K**VN\_ **V**DEYGG  
B GEEKAAV\_ **L**ALWD\_ **K**VN\_ **E**EEYGG  
C PADKTNV**K**AA\_ **W**G\_ **K**VGAHAGEYGA  
D AADKTNV**K**AA\_ **W**S\_ **K**VGGHAGEYGA  
E AA\_ **T**NV**K**TA\_ **W**SS**K**VGGHAPA\_ **A**

Once a gap, always a gap

# Probleme progressiver MSA-Verfahren

Angelehnt: Cedric Notredame, 2001

- 1: MAYFIELD THE LAST FAT RER
- 2: MAYFIELD THE FAST RAT
- 3: MAYLEENE IS A FAT RAT
- 4: MAYROONI THE LAST BIG RAT



```
MAYFIELD THE LAST FAT RER
MAYFIELD THE FAST RAT ____
MAYLEENE IS _ _ A _ FAT RAT
MAYROONI THE LAST BIG RAT
```

Besser:

```
MAYFIELD THE LAST FAT RER
MAYFIELD THE FAST ____ RAT
MAYLEENE IS _ _ A _ FAT RAT
MAYROONI THE LAST BIG RAT
```

# Verbesserungen

---

- Individuelle Scores für das **Öffnen eines Gaps** in Abhängigkeit der Umgebung, Abstand zu anderen Gaps, Länge der Sequenz, ...
- Verwendung **unterschiedlicher Substitutionsmatrizen**, je nachdem wie hoch man schon im Baum ist
  - Denn damit steigt der evolutionäre Abstand, und PAM-X bzw. BLOSUM-X Matrizen werden nach dem geschätzten Abstand gewählt

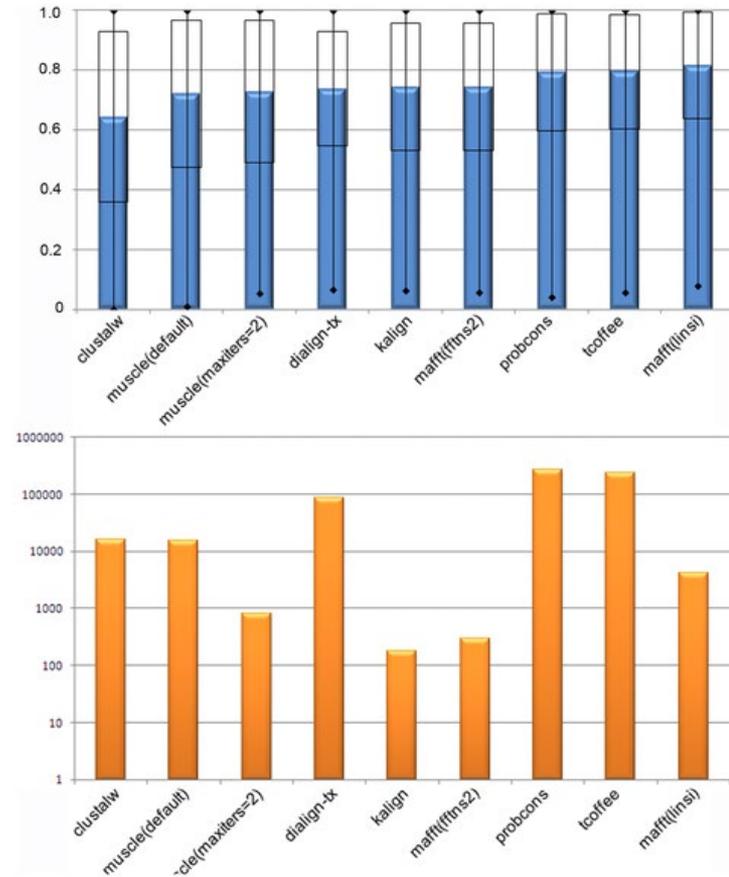
# Iterative Verfahren

---

- CLUSTAL W ist „greedy“
  - Ergebnis abhängig von der Struktur des Baumes
  - Der Guide Tree kann aber (evolutionär) falsch sein
  - Was am **Anfang schief läuft, ist besonders schlimm**
    - Alignments werden nie mehr korrigiert, sondern nur noch „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

# Stand der Technik

- Gold Standard wird über **Alignment der 3D-Strukturen** definiert
  - Strukturen aber für viele Proteine nicht bekannt
- Stark unterschiedliche Laufzeiten
  - Mehr als **hundertfache** Unterschiede
  - MSA ist komplex
  - 218 MSA,  $k_{\text{total}}=17892$



Thompson JD, Linard B, Lecompte O, Poch O (2011) A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLoS ONE 6(3)

# Inhalt dieser Vorlesung

---

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

# Suche mit MSA

---

- Erinnerung: Erzeugung von Proteinfamilien
  - Starte mit Proteinen gleicher/ähnlicher Funktion
  - Finde das Gemeinsame durch MSA
  - Suche mit dem MSA nach weiteren Vertretern
  - Modifiziere Familie entsprechend
  - Iteriere, bis Zufriedenheit eintritt
- Wie sucht man mit einem MSA?
  - Wir müssen entscheiden, wie gut eine gegebene Sequenz  $S$  zu einem gegebenen MSA  $M$  passt
  - Verschiedene Möglichkeiten: RegExp, Profile, Profile-HMM

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

Quelle: [Kro98]

- Vorteil: Schnell berechnet, **schnelles Matching**
- Nachteile
  - Unklare Behandlung von INDEL (wann fügt man \* ein?)
  - Keine Berücksichtigung der Häufigkeit von Zeichen (solange >0)
  - 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]

T	G	C	T	A	G	G	Eher schlecht
---	---	---	---	---	---	---	---------------

A	C	A	C	A	T	C	Eher gut
---	---	---	---	---	---	---	----------

# Inhalt dieser Vorlesung

---

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

# Variante 2: Profile

- Definition

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

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

- Beispiel

S <sub>1</sub>	A	G	C	-	A
S <sub>2</sub>	A	G	A	G	A
S <sub>3</sub>	T	C	C	G	-
S <sub>4</sub>	C	G	-	T	C
A	0.50	0	0.25	0	0.50
G	0	0.75	0	0.5	0
C	0.25	0.25	0.50	0	0.25
T	0.25	0	0	0.25	0
-	0	0	0.25	0.25	0.25

# Profile und Sequenzen

---

- Mit dem Profil P eines MSA M kann man bewerten, wie gut eine Sequenz S mit M matched
- Setzt **Alignment von S mit M** voraus
  - Welche Zeichen der Sequenz sollen mit welchen Spalten des MSA verglichen werden?
- Wie immer braucht man zwei Dinge
  - Methode zur Bewertung eines **konkreten Alignments** von S und P
  - Methode zum Finden des **besten Alignments** zwischen S und P

# 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

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

# Beispiel: S = AAGGC

Profil P

	1	2	3	4	5
A	0.50	0	0.25	0	0.50
G	0	0.75	0	0.5	0
C	0.25	0.25	0.50	0	0.25
T	0.25	0	0	0.25	0
-	0	0	0.25	0.25	0.25

Substitutionsmatrix m

	A	G	C	T	-
A	2	-1	-3	-1	-2
G		2	-1	-1	-2
C			2	-1	-2
T				2	-2
-					0

Alignment A

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

$$\begin{aligned}
 s(A) &= (2*0.5 + -1*0 + -3*0.25 + -1*0.25) + (-2) + \\
 &\quad (-1*0 + 2*0.75 + -1*0.25 + -1*0) + \\
 &\quad (-2*0.25 + -2*0 + -2*0.50 + -2*0.25) + \\
 &\quad (-1*0 + 2*0.5 + -1*0 + -1*0.25) + \\
 &\quad (-3*0.50 + -2*0 + 2*0.25 + -1*0) \\
 &= 0 - 2 + 1.25 - 2 + 0.75 - 1 \\
 &= -3
 \end{aligned}$$

# Optimale Profilalignments

---

- Theorem

*Gegeben eine Substitutionsmatrix  $m$ , Profil  $P$ , Sequenz  $S$*

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

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

- *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*

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

# Anwendung: PSI-BLAST

---

- Implementierung der iterativen Suchstrategie
  - Gegeben Suchsequenz S: Berechne Profil P
  - Durchsuche DB mit P (hier: ohne Gaps)
  - Bilde multiples Alignment aller Hits
  - Berechne daraus eine neues P
  - Iteriere, bis Stoppkriterium erfüllt
- **Erhöhung der Sensitivität** gegenüber einfachem BLAST
  - PSI-BLAST findet auch weiter entfernte Homologien
  - Zwitter zwischen Patternmatching und Homologiesuche
  - Gefahr der „drift“

# Inhalt dieser Vorlesung

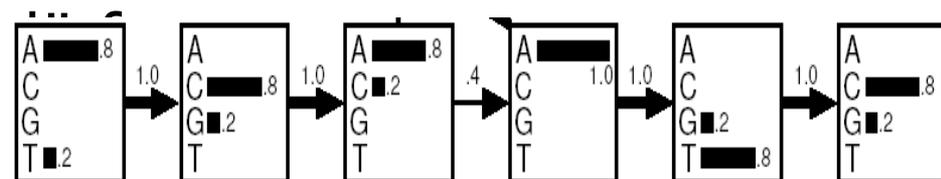
---

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

# Grundidee

- Wir modellieren ein MSA M als **Sequenz von Spalten**
- Eine Spalte wird **ein Zustand**, der prinzipiell alle Zeichen ausgeben kann – Wsk ergibt sich aus den relativen Häufigkeiten in der Spalte
- Beim Berechnen kennen wir den Zustand nicht - also brauchen wir **ein HMM**
- Das HMM ist ein mehr oder weniger gutes **Modell für eine Suchsequenz**
  - Das berechnet uns Viterbi / Forward
- Allerdings müssen wir auch Spalten bzw. Zeichen überspringen
  - Spezielle Zustände im 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



# Erste Idee

---

- Jede „hinreichend volle“ Spalte wird ein **Match-Zustand**
  - „Hinreichend voll“: Schwellwert für die Anzahl an INDELS
  - **Emissionswsken**: Häufigkeiten der Zeichen in der Spalte
- „Halbvolle, nicht vollkommen leere“ Spalten darf man überspringen: **Insertion-Zustand**
  - Mehrere INS Zustände hintereinander können zu Blöcken zusammengefasst werden
  - **Emissionswsken**: Häufigkeiten der Zeichen im Block
- „Fast-leere“ Spalten lassen wir gleich ganz weg
- **Übergangs-Wsk** ergeben sich aus der Sequenz der Spalten und Insert Zustände

# 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

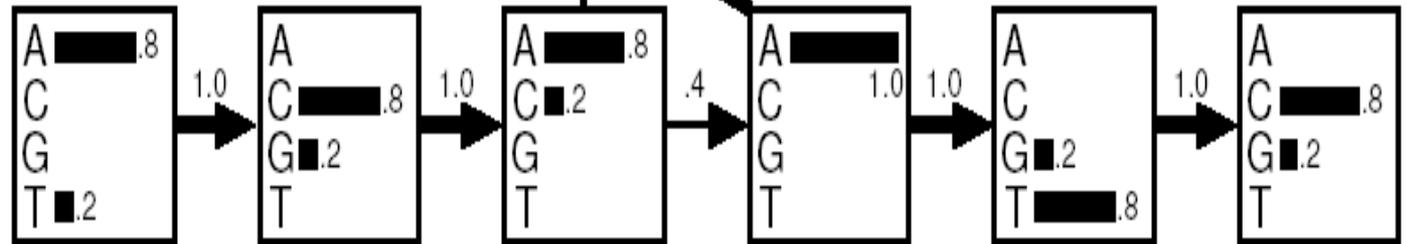
Verteilung in den Spalten 4-6

In 3/5 Fällen geht es nach I mit M weiter

4-6

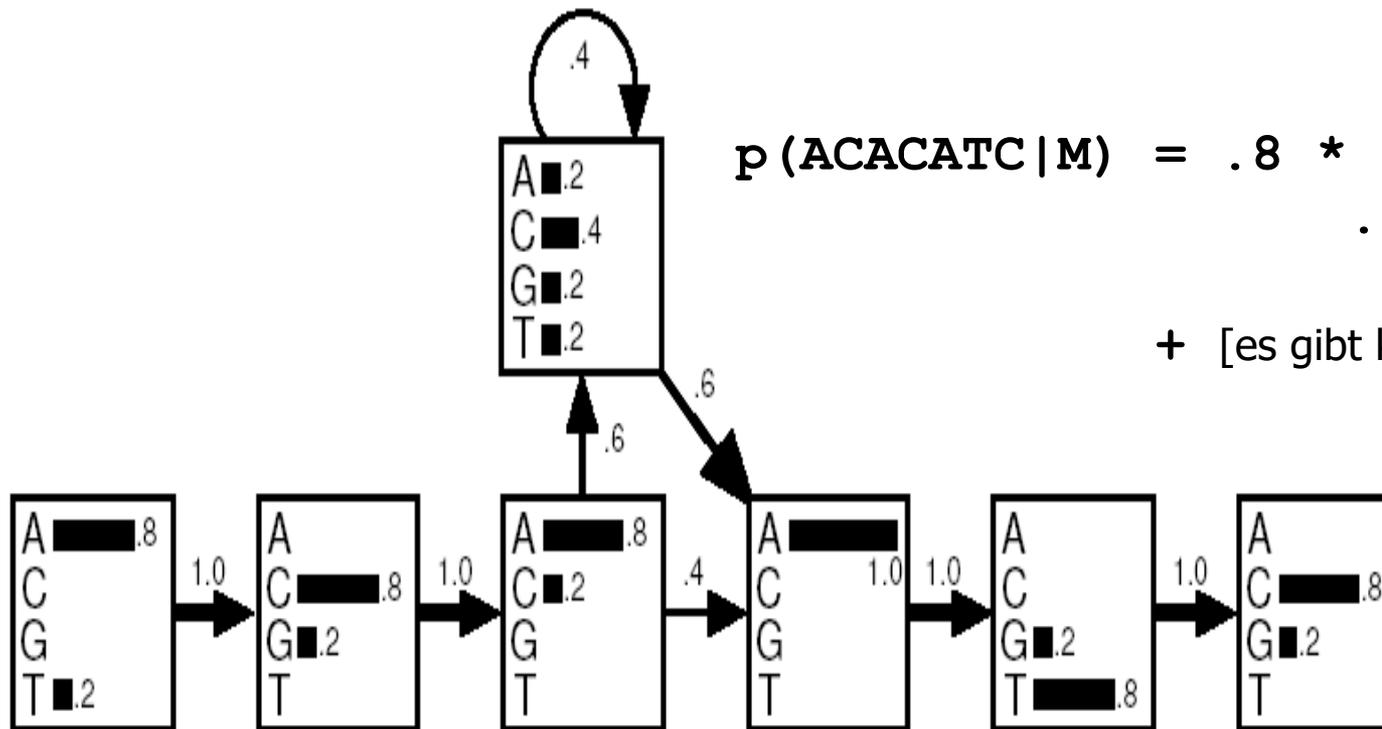
Matches

[Kro98]



# Scoring

- Berechnung von  $p(S|M)$ 
  - M: Modell des MSA, S die zu analysierende Sequen
  - S=ACACATC



$$p(ACACATC | M) = .8 * 1 * .8 * 1 * .8 * .6 * .4 * .6 * 1 * 1 * .8 * 1 * .8$$

+ [es gibt keinen anderen Pfad]

# Beispielscores (Bester Pfad)

	<b>Sequenz</b>	<b>Wsk (%)</b>
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

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

# Länge einer Sequenz

	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 durch **Log-Odds Score**
  - Nullmodell: Wsk, dass die Sequenz zufällig aus gleichverteilten Wsk erzeugt wurde

# Log Odds Scores

---

	<b>Sequenz</b>	<b>Wsk (%)</b>	<b>Log-odds</b>
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

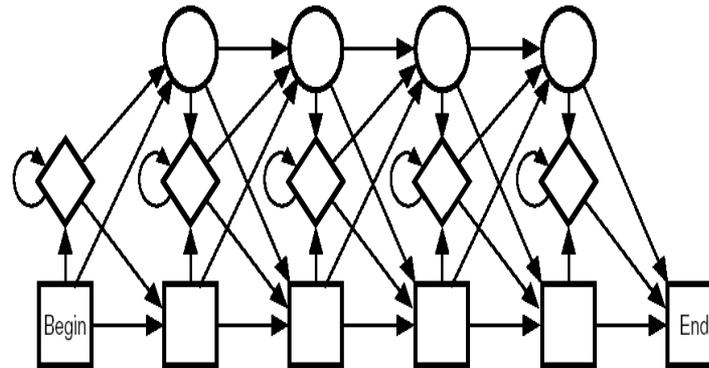
# Profil-HMMs

---

- Das eben angewandte Verfahren zur Umwandlung einer MSA in ein HMM trifft einige erratische Entscheidungen
  - Warum keine Insertions an anderen Stellen erlauben? Die sollten zwar bestraft werden, aber möglich sein
  - Es gibt keine Deletions
- Profil-HMMs: **HMM mit spezieller Struktur**
  - Feste, repetitive Zustands-Struktur
  - Wsk der INDELS wird durch **Übergangs-Wsk** ausgedrückt

# Struktur

---

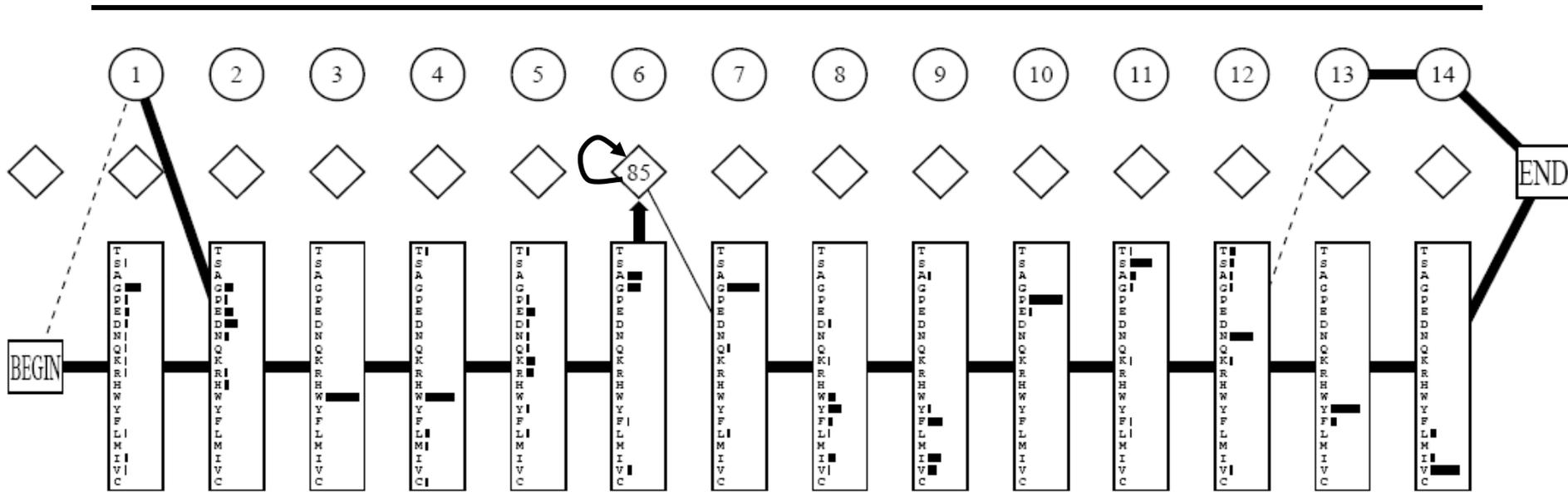


- Rechtecke: Match-Zustände
  - Symbolisieren relativ volle Spalten (Typisch: >50%)
- Rauten: Insertion-Zustände
  - Symbolisieren **Spalten/Bereiche mit vielen Gaps**
- Kreise: Deletion-Zustände
  - Überspringen genau 1 Match-Zustand, aber können verkettet sein
  - „**Silent States**“: Emittieren keine Zeichen

# Beispiel

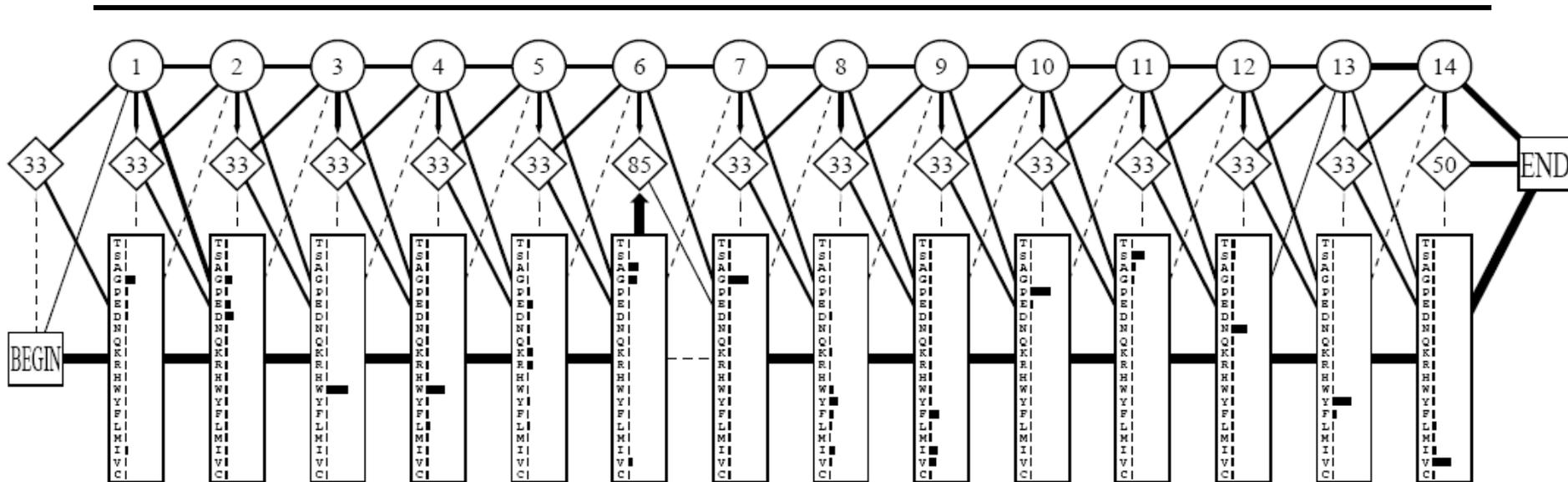
```
GGWWRRG d y . g g k k q L W F P S N Y V
IGWLNNG dyne t t g e r r G D F P G T Y V
PNWWE G q l . . n n r r G I F P S N Y V
DEW W Q A r r . . d e q q i G I V P S N Y V
GEW W K A q s . . t g q q e G F I P F N F V
GDW W L A r s . . s g q q t G Y I P S N Y V
G D W W D A e l . . k g r r r G K V P S N Y L
- D W W E A r s l s s g h r r G Y V P S N Y V
G D W W Y A r s l i t n s e G Y I P S T Y V
GEW W K A r s l a t r k e G Y I P S N Y V
GDW W L A r s l v t g r e G Y V P S N F V
GEW W K A k s l s s k r e G F I P S N Y V
GEW C E A q t . k n g q . G W V P S N Y I
SDW W R V v n l t t r q e G L I P L N F V
LPW W R A r d . k n g q e G Y I P S N Y I
RDW W E F r s k t v y t p G Y Y E S G Y V
EHW W K V k d . a l g n v G Y I P S N Y V
IHW W R V q d . r n g h e G Y V P S S Y L
KDW W K V e v . . n d r q G F V P A A Y V
VGW M P G l n e r t r q r G D F P G T Y V
PDW W E G e l . . n g q r G V F P A S Y V
ENW W N G e i . . g n r k G I F P A T Y V
EEW L E G e c . . k g k v G I F P K V F V
GGW W K G d y . g t r i q Q Y F P S N Y V
DGW W R G s y . . n g q v G W F P S N Y V
QGW W R G e i . . y g r v G W F P A N Y V
GRW W K A r r . . a n g g e t G I I P S N Y V
GGW T Q G e l . . k s g q k G W A P T N Y L
GDW W E A r s n . t g g e n G Y I P S N Y V
NDW W T G r t . . n g k e G I F P A N Y V
```

# Das Profil-HMM dazu



- Schlecht konservierter Block wird ein einziger INS Zustand
  - 85% Wsk für Übergang zum selben Zustand
  - Schätzen der Parameter: Später
- Offensichtliches Problem: **Overfitting**
  - Deletions sind praktisch überall verboten

# Profil-HMM mit Pseudo-Counts



- **Smoothing** mit Pseudo-Count von 1
- Sehr wichtig, da Sequenzfamilien selten 1000nde Sequenzen umfassen
  - Und das Profil-HMM sehr viele Zustände hat

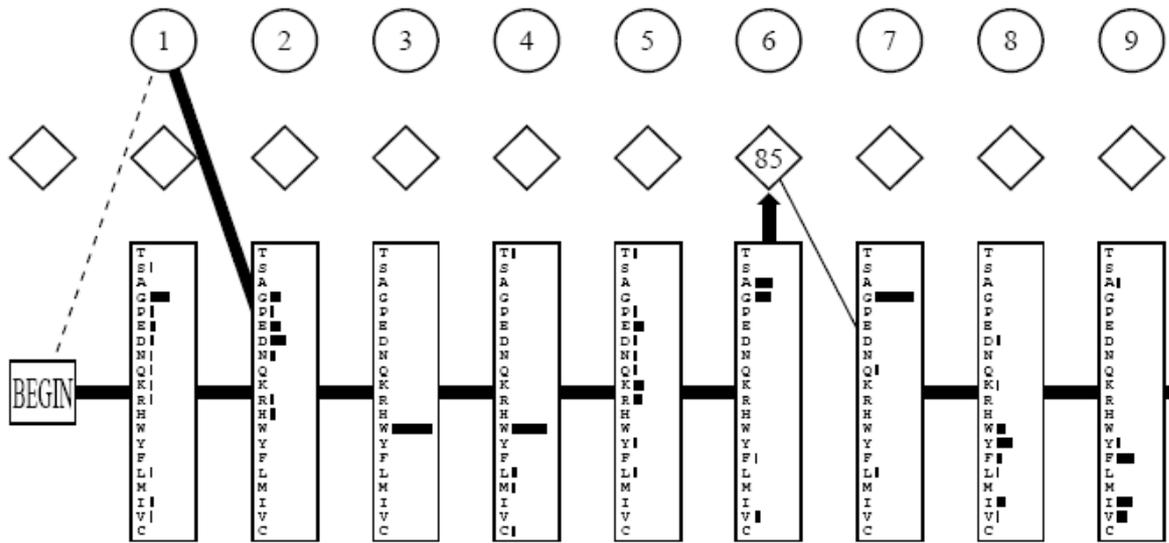
# Lernen eines Profile-HMM

---

- Wir haben den (glücklichen) Fall 1 des Trainingsproblems für HMM
  - Wir benutzen die Sequenzen des MSA als Trainingsdaten
  - Alle Sequenzen des MSA sind im MSA aligniert
  - Damit **wissen wir zu jeder Sequenz genau die Zustandsfolge**
  - Also: Maximum Likelihood Schätzung (hier 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	L	W	F	P	P	S	N	Y	V	
I	G	W	L	N	G	g	q	l	.	n	n	r	r	G	D	F	P	S	N	Y	V	
P	N	W	W	E	G	r	r	.	.	d	e	q	q	G	I	F	P	S	N	Y	V	
D	E	W	W	Q	A	r	r	.	.	s	g	g	q	G	I	V	P	S	N	-	-	
G	D	W	W	L	A	r	r	.	.	t	s	g	g	G	Y	I	P	S	N	F	V	
G	D	W	W	D	A	r	r	.	.	k	g	g	r	G	K	V	P	S	N	Y	L	
-	D	W	W	E	A	r	r	.	.	s	l	s	g	G	Y	V	P	S	N	Y	V	
G	D	W	W	Y	A	r	r	.	.	l	i	t	n	G	Y	I	P	S	T	Y	V	
G	G	E	W	W	K	A	r	r	.	s	l	a	t	G	Y	I	P	S	N	Y	V	
G	G	E	W	W	L	A	r	r	.	s	l	v	s	G	Y	I	P	S	N	F	V	
G	G	E	W	W	L	A	k	r	.	s	l	v	s	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	r	v	.	n	.	l	t	t	G	L	I	P	L	N	F	V	
L	P	W	W	R	A	r	r	.	.	k	t	n	g	G	Y	I	P	S	N	Y	I	
R	D	W	W	E	F	r	r	.	.	s	k	.	t	G	Y	I	P	S	S	Y	V	
E	H	W	W	K	V	k	d	.	.	r	.	a	l	G	Y	I	P	S	S	Y	V	
I	H	D	W	W	R	V	q	.	.	e	.	.	.	G	Y	I	P	S	S	Y	V	
V	G	W	W	M	P	V	g	.	.	e	.	.	.	G	D	F	P	P	A	T	V	
P	D	W	W	E	G	e	l	.	.	n	g	g	q	G	V	F	P	A	S	Y	V	
E	N	W	W	N	G	e	i	.	.	g	n	r	k	G	I	F	P	P	K	V	V	
E	E	W	W	L	E	G	e	.	.	k	g	k	v	G	I	F	P	P	S	N	Y	V
G	G	W	W	K	G	d	y	.	.	g	t	r	r	G	Q	W	F	P	S	N	Y	V
D	G	W	W	R	G	s	e	.	.	y	g	g	g	G	W	F	P	P	A	N	Y	V
G	G	W	W	T	O	G	e	.	.	a	n	g	g	G	I	I	P	S	N	Y	V	
G	D	W	W	E	A	r	r	.	.	k	s	g	g	G	W	A	P	T	N	Y	L	
N	D	W	W	T	G	r	t	.	.	n	.	g	e	G	Y	I	P	S	N	Y	V	

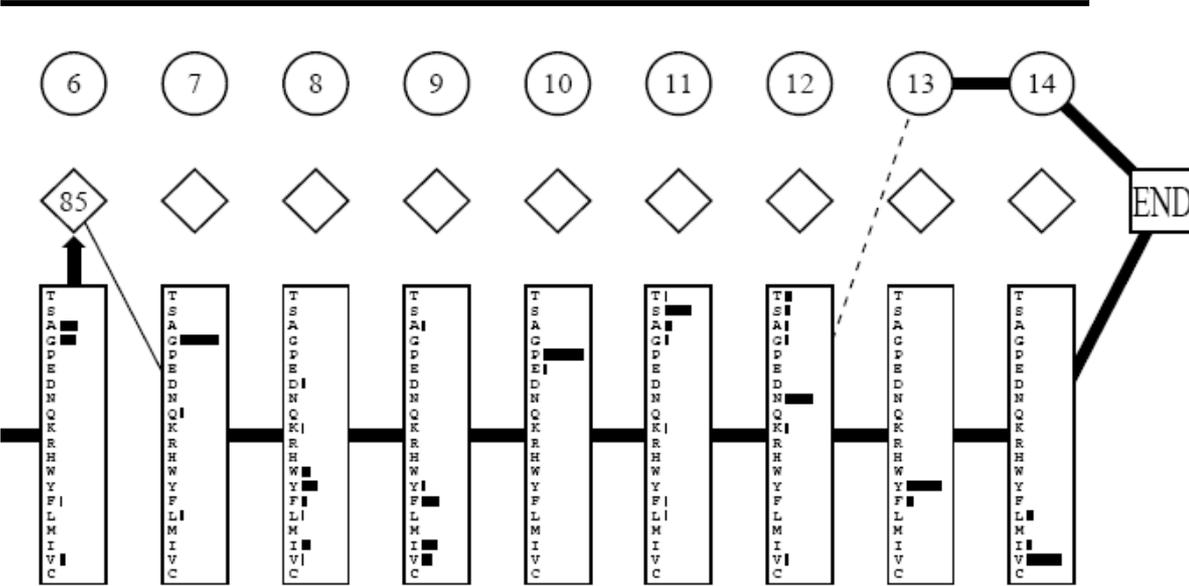
- Erste Spalte: Gap=> Wsk=1/30 für Start-DEL, 29/30 für Start-Match1
- Keine Gaps in Spalten 2-6: Wsk 1 für Übergänge zwischen Matches
- Nach Match6 kommt mit 100% Wsk eine Insertion, keine Sequenz hat nur Leerzeichen zwischen den beiden „guten“ Blöcken
- Zählen aller Übergänge INS-INS und INS-Match7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

# Profile-HMM erklärt

```

GGWLNRRGdyne.gttgkkrqLGGDFFPSSNTYYVV
PNWWEQGAqrll...nngkrrGGIIFPSSNYV
DEWVWQAARrrs...dneqqqGGIIFPSSNK-
GDWVWLARsrl...sttqqqGGYIIPSSNFV
GGDWWDAARsl...skqqqrrGGKIVPSSNYLV
-DWWEAARslsllsttngghrGGYVPPSSNYVV
GGDWWYKARslslattnngkrrGGYIIPSSNYVV
GGDWWLARslsvttngkrrGGYVPPSSNFVV
GGEWVKAKstlsskngrrq...GGFVPPSSNYVI
GGEWCEAqt.tkknggrrq...GGWVPPSSNYIV
SSDWWRAVr.vnl.tttnggrrq...GGYIIPSSNFIV
LRDWWRFrsk.kttngvyygrrq...GGYIIPSSNYIV
IHKWWRVqel.vner.nngkrrq...GGYVPPSSNYLV
IKVDWWMKPGel...nggkrrq...GGFVPPSSATYVV
PENWVNGEgei...ngkrrq...GGIIFPSSNYVV
EEWLEGGedyy...gkngkrrq...GGIIFPSSNYVV
DDGWWRKGGsry...yngkrrq...GGIIFPSSNYVV
OQGWWRGAGrrl...kngkrrq...GGIIFPSSNYVV
GGGWWTQAGrrt...nngkrrq...GGIIFPSSNYLV
GGDWWT...nngkrrq...GGIIFPSSNYVV
NDWWT...nngkrrq...GGIIFPSSNYVV

```



- 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

---

- Man kann alle Varianten anwenden
  - Viterbi: Insgesamt wahrscheinlichstes Alignment
  - Forward: Gesamtwahrscheinlichkeit der Sequenz
  - Forward/Backward: Lokale Wahrscheinlichkeiten jeder Ausgabe

# Selbstkontrolle

---

- Welche Struktur hat ein Profil-HMM?
- Was unterscheidet ein Profil-HMM und ein Profil (auch PSWM) bei der Anwendung in der Suche konzeptionell?
- Wie werden schlecht konservierte Blöcke in Profil-HMMS repräsentiert? Wo gibt es hier Design-Freiheiten?
- Warum löst man MSA praktisch nur mit Heuristiken? Warum nennen wir diese Verfahren überhaupt Heuristiken?
- Komplexität von Clustal-W mit hierarchischem Clustering?