Algorithmische Bioinformatik

CLUSTAL W

Profilalignment

Profile HMMs

Ulf Leser

Wissensmanagement in der Bioinformatik

Ziele der Vorlesung

- Die klassische Heuristik für MSA kennenlernen
- Suche mit MSAs – Verhältnis HMM zu Pattern-Matching verstehen
  - HMMs als weiteres Verfahren zum approximativen Stringmatching

Inhalt dieser Vorlesung

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

CLUSTAL W

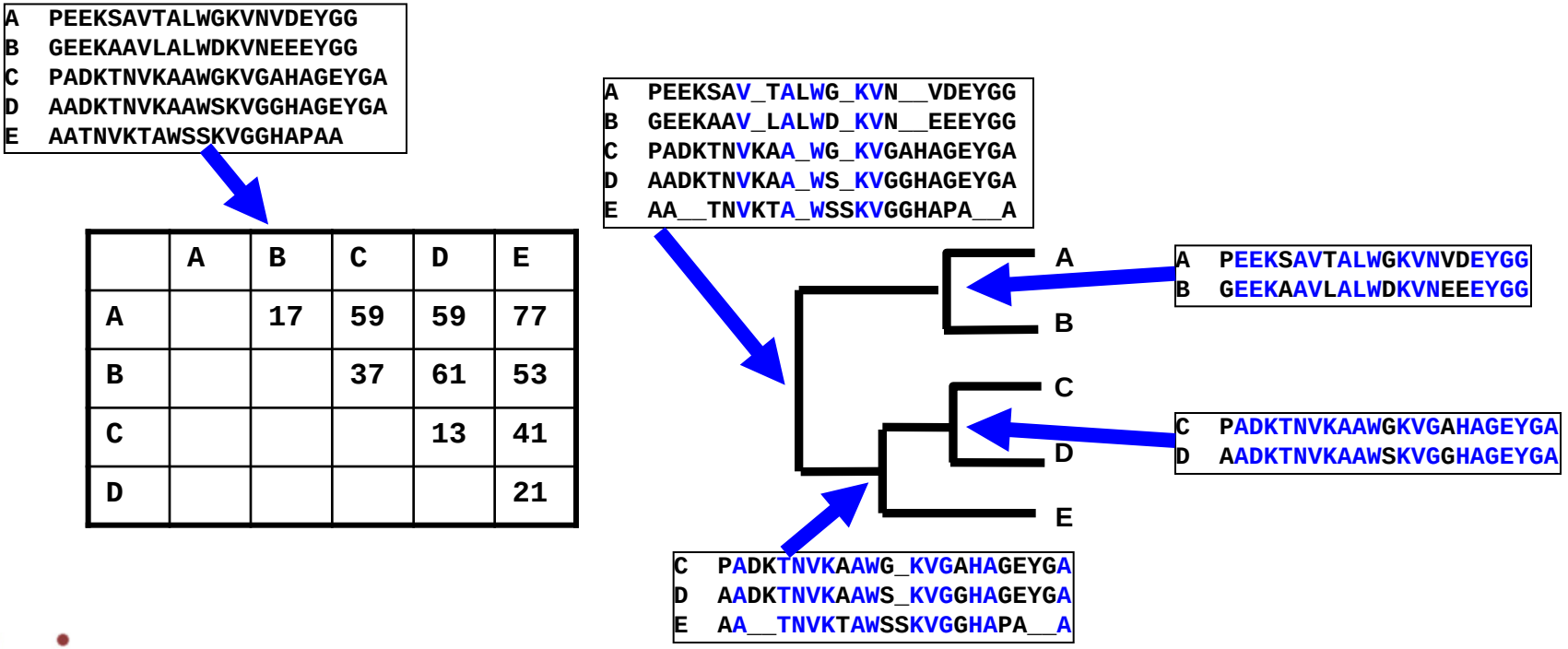
- Greedy-Variante von „MSA mit phylogenetischem Baum“
- Lange Zeit das Standardprogramm für MSA
  - Higgins, D. G. and Sharp, P. M. (1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer." *Gene* **73**(1): 237-44.
  - Thompson, J. D., Higgins, D. G. and Gibson, T. 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 Alignmentscores
  - Konstruiere einen „Guide Tree“
  - Berechne und verschmelze Teil-MSA gemäß dem Guide Tree

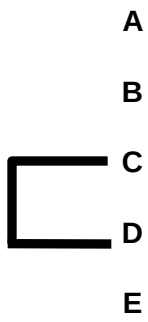


Schritt 1 und 2

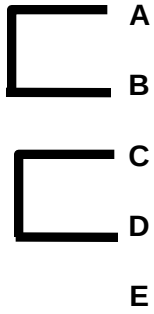
- Schritt 1: Berechnen der Ähnlichkeitsmatrix M
  - $O(k^2)$ paarweise Alignments
- Schritt 2: Hierarchisches Clustering (ursprünglich)
  - Wähle Zelle (i,j) mit kleinstem Abstand aus Matrix M
    - Das ist das erste Paar
  - Erzeuge M‘: Lösche die Sequenzen i und j aus M und füge neue Spalte/Zeile (ij)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

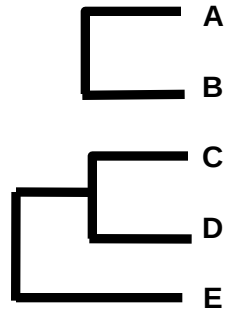
	A	B	C	D	E
A		17	59	59	77
B			37	61	53
C				13	41
D					21

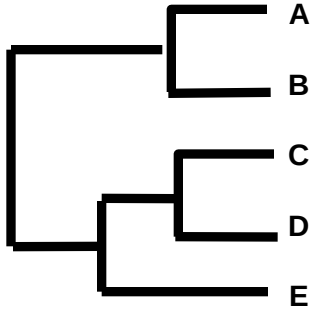


	A	B	E	CD
A		17	77	59
B			53	49
E				31



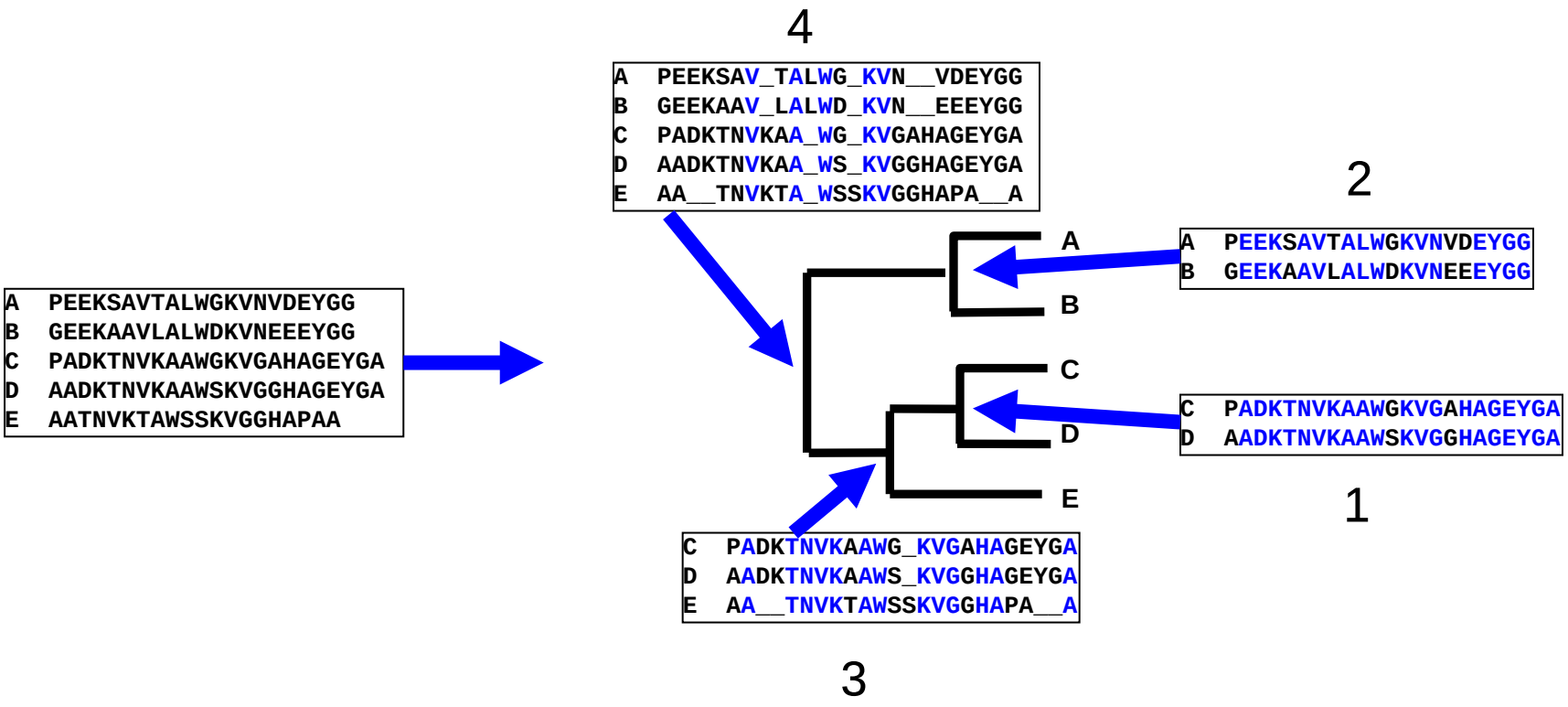
	E	CD	AB
E		31	65
CD			54





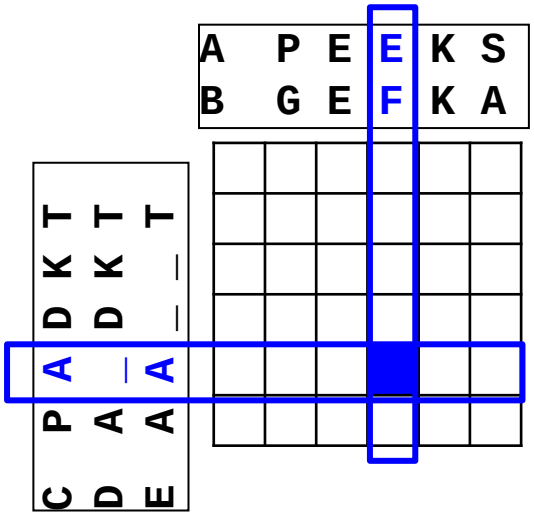
Schritt 3: Progressive MSA Generierung

- 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 Alignement 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)



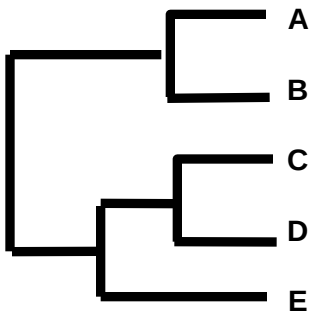
Score des Spaltenvergleichs:

```
(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



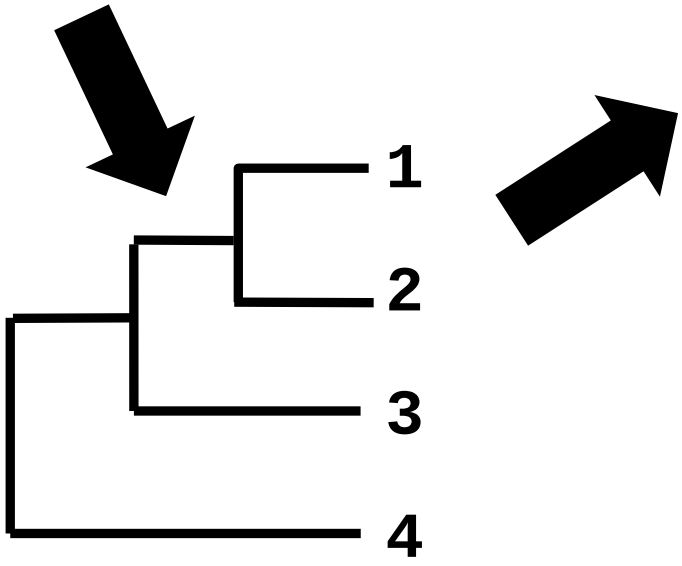

```
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA

A   PEEKSAVTALWGKVNVDEYGG
B   GEEKAAVLALWDKVNEEEYGG

C   PADKTNVKAAWG_KVGAHAGEYGA
D   AADKTNVKAAWS_KVGGHAGEYGA
E   AA__TNVKTAWSSKVGGHAPA__A

A   PEEKSAV_TALWG_KVN__VDEYGG
B   GEEKAAV_LALWD_KVN__EEEYGG
C   PADKTNVKAA_WG_KVGAHAGEYGA
D   AADKTNVKAA_WS_KVGGHAGEYGA
E   AA__TNVKTA_WSSKVGGHAPA__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 RAT
```




```
MAYFIELD THE LAST FAT RER
MAYFIELD THE FAST RAT ___
MAYLEENE IS_ _A__ FAT RAT
MAYROONI THE LAST ___ RAT
```

Besser:

```
MAYFIELD THE LAST FAT RER
MAYFIELD THE FAST ___ RAT
MAYLEENE IS_ _A__ FAT RAT
MAYROONI THE LAST ___ RAT
```

Verbesserungen

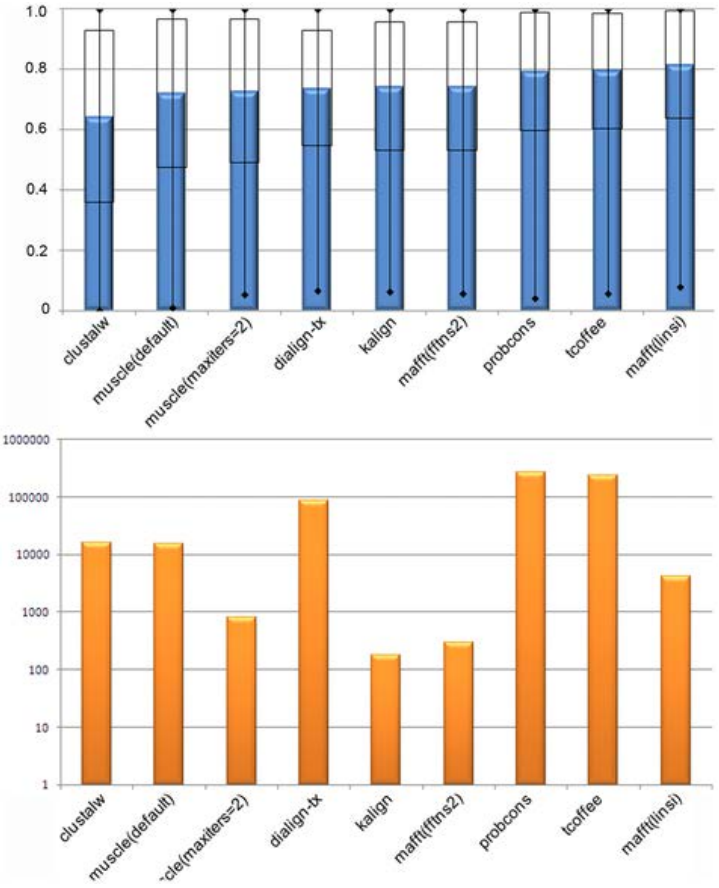
- Individuelle Scores für das Öffnen eines Gaps in Abhängigkeit der Umgebung, Abstand zu anderen Gaps, Länge der Sequenz, …
- Verwendung unterschiedlicher Substitutionsmatrizen, je nachdem wie hoch man schon im Baum ist
  - Denn damit steigt der evolutionäre Abstand, und PAM-X bzw. BLOSUM-X Matrizen 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_{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*|Σ'|*
  - *(i,j) enthält die relative Häufigkeit des Zeichens j in der Spalte i*
- Beispiel

S_1	A	G	C	-	A
S_2	A	G	A	G	A
S_3	T	C	C	G	-
S_4	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*

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

Beispiel: S = AAGGC

Profil P

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

```
s(A) = (2*0.5   + -1*0   + -3*0.25 + -1*0.25)    +
       (-2) +
       (-1*0    + 2*0.75 + -1*0.25 + -1*0)     +
       (-2*0.25 + -2*0   + -2*0.50 + -2*0.25) +
       (-1*0    + 2*0.5  + -1*0    + -1*0.25)  +
       (-3*0.50 + -2*0   + 2*0.25  + -1*0)
     = 0 – 2 + 1.25 – 2 + 0.75 – 1
     = -3
```

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*

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

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“

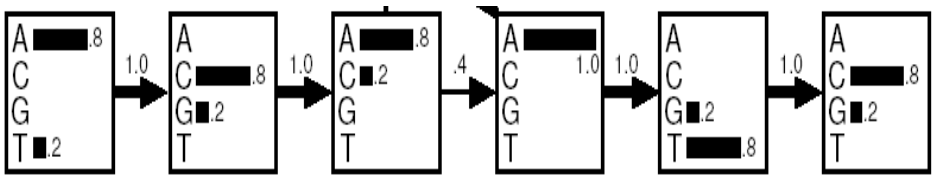
Inhalt dieser Vorlesung

- 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
  - 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 dürfen
  - 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 Zustande 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

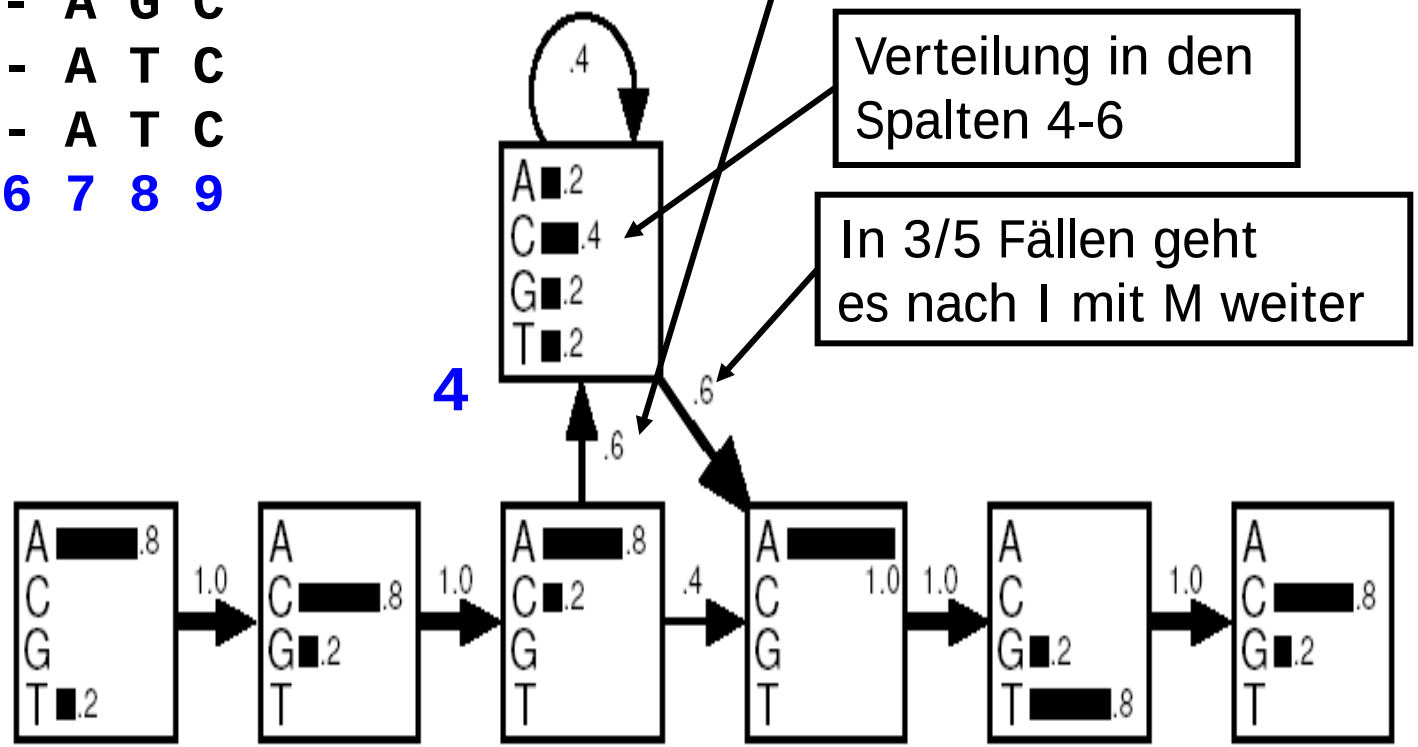
A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C
1	2	3	4	5	6	7	8	9

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



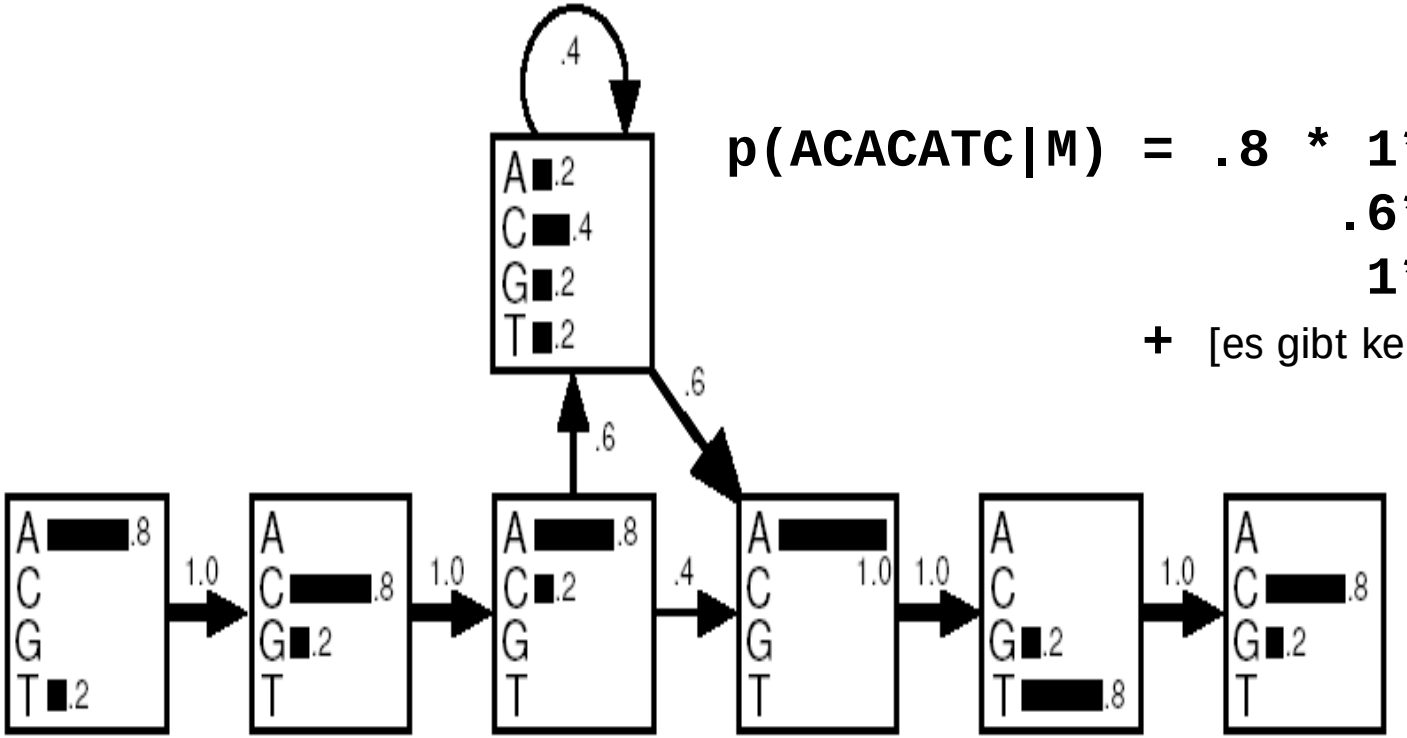
Matches

[Kro98]

1 2 3 7 8 9

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)

	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

```
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

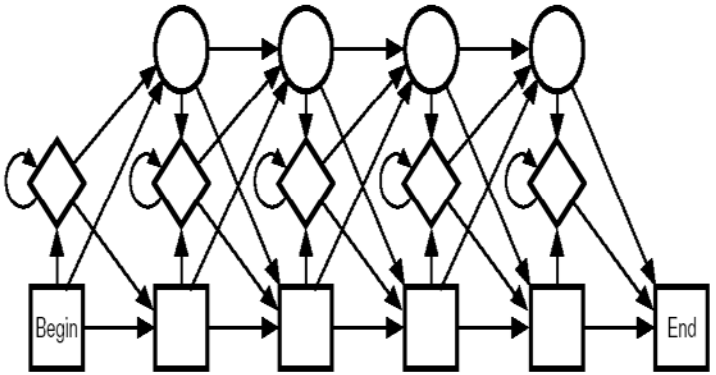
Log Odds Scores

	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

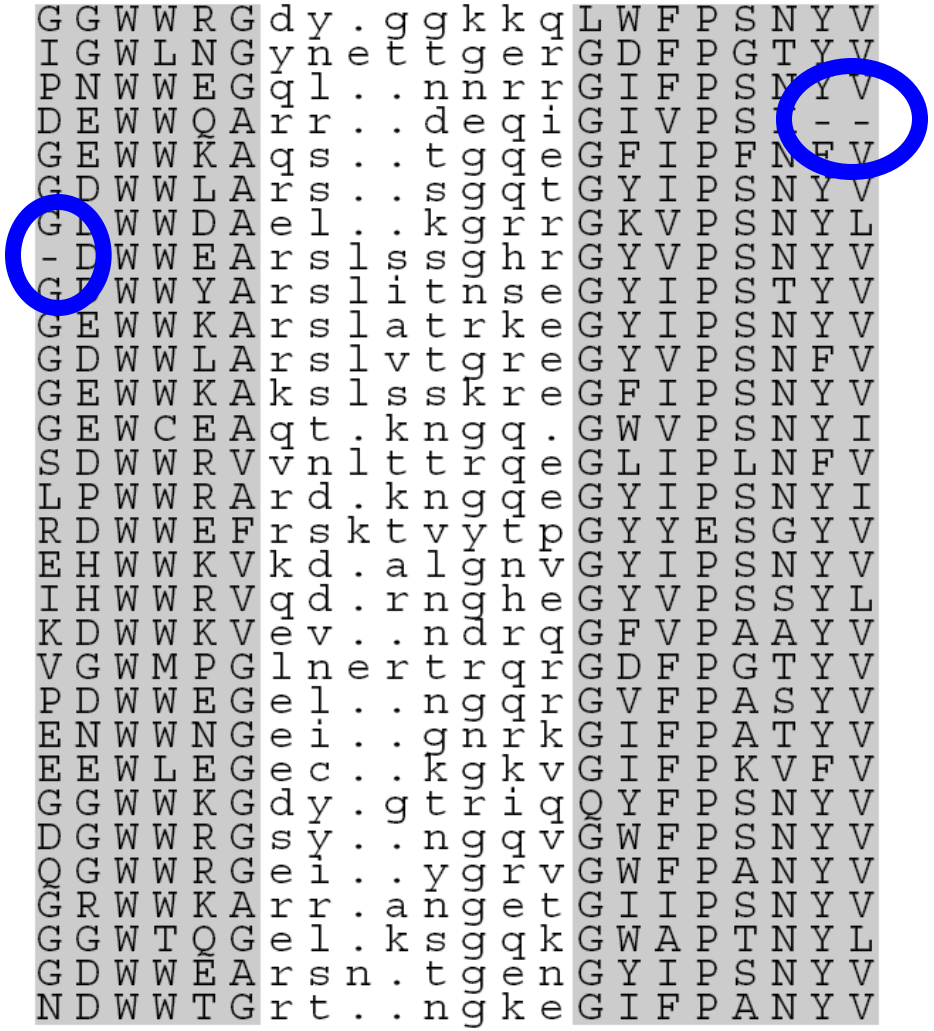
- Das eben angewandte Verfahren zur Umwandlung einer MSA in ein HMM trifft einige erratische Entscheidungen
  - Warum keine Insertions an anderen Stellen erlauben? Die sollten zwar bestraft werden, aber möglich sein
  - Warum keine Deletions?
- Profil-HMMs: HMM mit spezieller Struktur
  - Feste, repetitive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

Struktur

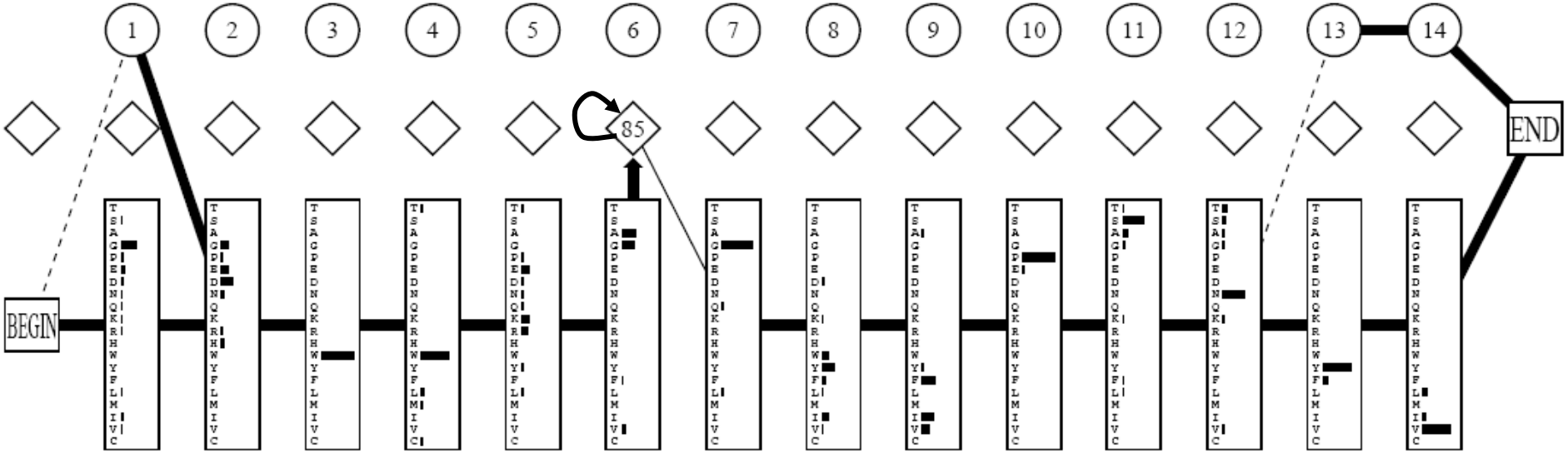


- Rechtecke: Match-Zustände
  - Symbolisieren relativ volle Spalten (Typisch: >50%)
- Rauten: Insertion-Zustände
  - Symbolisieren Spalten/Bereiche mit vielen Gaps
- Kreise: Deletion-Zustände
  - Überspringen genau 1 Match-Zustand, aber können verkettet sein
  - „Silent States“: Emittieren keine Zeichen

Beispiel

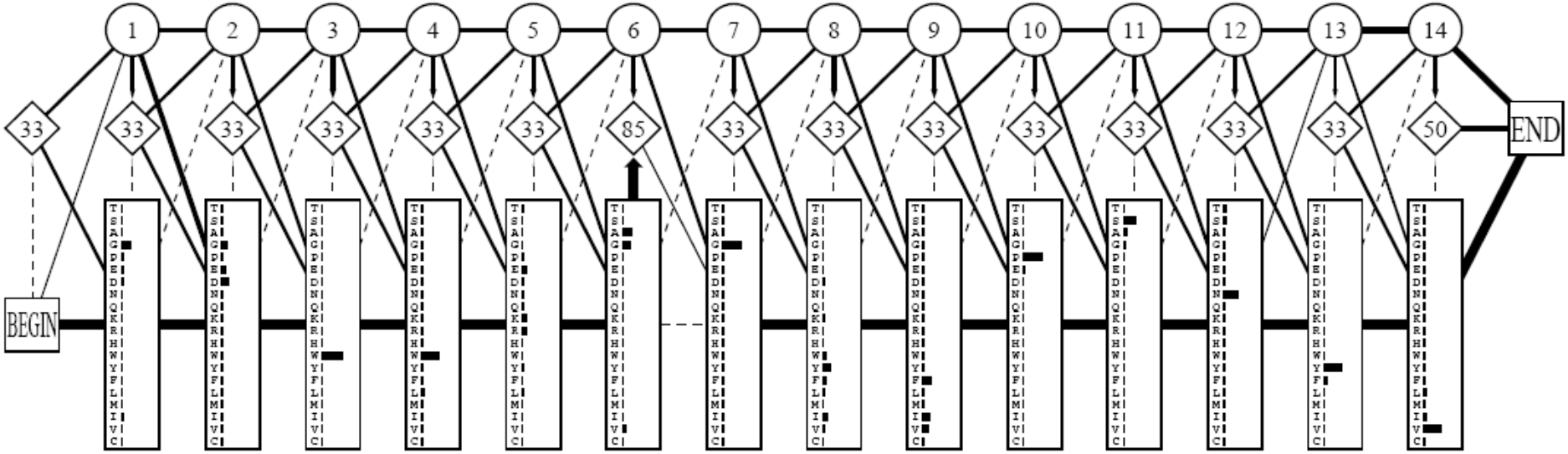

G G W W R G d y . g g k k q L W F P S N Y V
I G W L N G y n e t t g e r G D F P G T Y V
P N W W E G q l . . n n r r G I F P S N Y V
D E W W Q A r r . . d e q i G I V P S I - -
G E W W K A q s . . t g q e G F I P F N F V
G D W W L A r s . . s g q t G Y I P S N Y V
G L W W D A e l . . k g r r G K V P S N Y L
- D W W E A r s l s s g h r G Y V P S N Y V
G I W W Y A r s l i t n s e G Y I P S T Y V
G E W W K A r s l a t r k e G Y I P S N Y V
G D W W L A r s l v t g r e G Y V P S N F V
G E W W K A k s l s s k r e G F I P S N Y V
G E W C E A q t . k n g q . G W V P S N Y I
S D W W R V v n l t t r q e G L I P L N F V
L P W W R A r d . k n g q e G Y I P S N Y I
R D W W E F r s k t v y t p G Y Y E S G Y V
E H W W K V k d . a l g n v G Y I P S N Y V
I H W W R V q d . r n g h e G Y V P S S Y L
K D W W K V e v . . n d r q G F V P A A Y V
V G W M P G l n e r t r q r G D F P G T Y V
P D W W E G e l . . n g q r G V F P A S Y V
E N W W N G e i . . g n r k G I F P A T Y V
E E W L E G e c . . k g k v G I F P K V F V
G G W W K G d y . g t r i q Q Y F P S N Y V
D G W W R G s y . . n g q v G W F P S N Y V
Q G W W R G e i . . y g r v G W F P A N Y V
G R W W K A r r . a n g e t G I I P S N Y V
G G W T Q G e l . k s g q k G W A P T N Y L
G D W W E A r s n . t g e n G Y I P S N Y V
N D W 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 Übergangs zum selben Zustand
  - Schätzen der Parameter: Später
- Offensichtliches Problem: Overfitting
  - Deletions sind praktisch überall „verboten“

Profil-HMM mit Pseudo-Counts



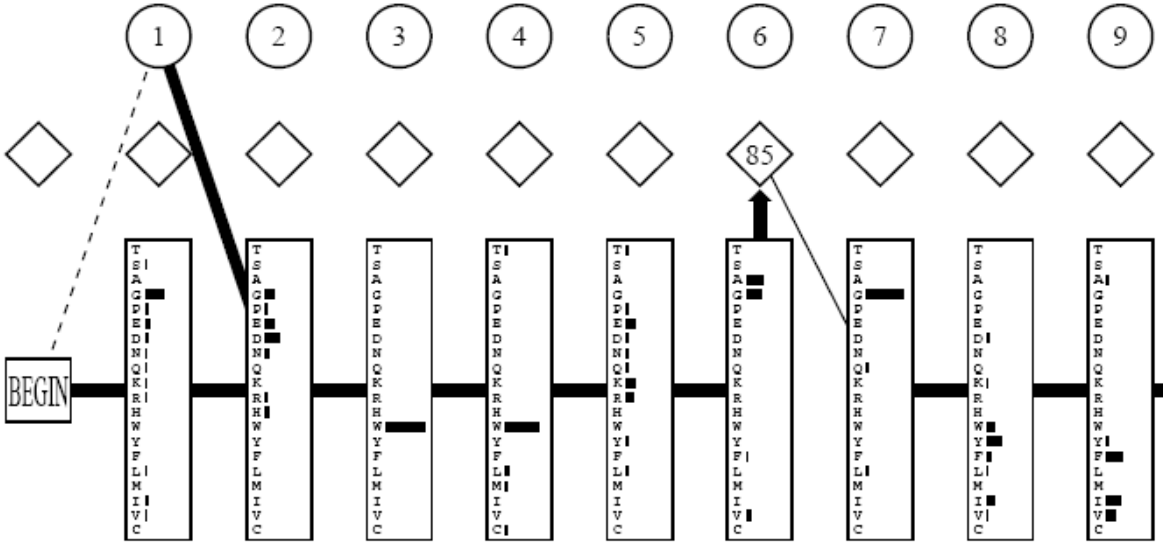
- Smoothing mit Pseudo-Count von 1
- Sehr wichtig, da Sequenzfamilien selten 1000nde Sequenzen umfassen
  - Und das Profil-HMM sehr viele Zustände hat

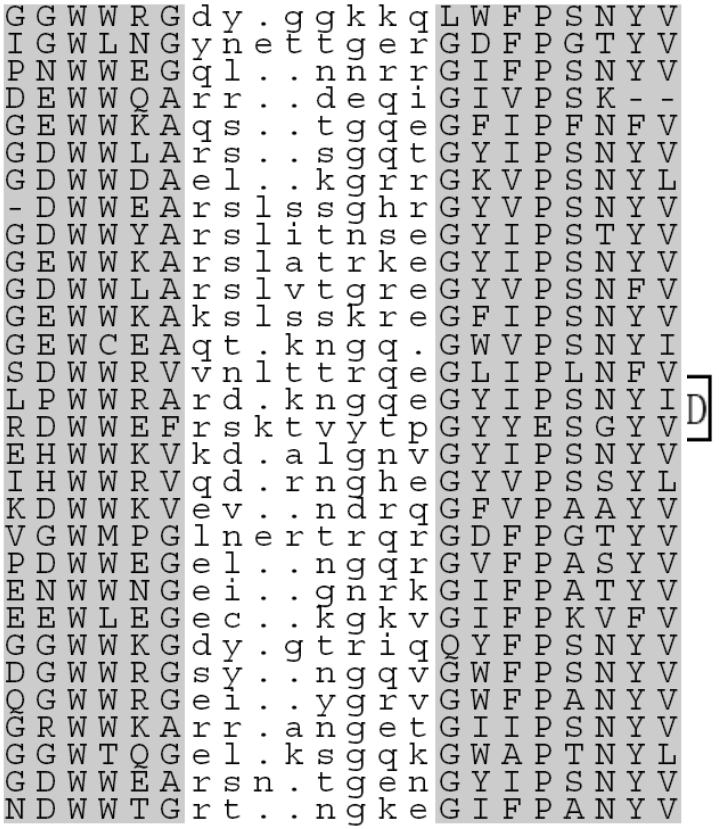
Lernen eines Profile-HMM

- Wir haben den (glücklichen) Fall 1 des Trainingsproblems für HMM
  - Wir benutzen die Sequenzen des MSA als Trainingsdaten
  - Alles Sequenzen des MSA sind im MSA aligniert
  - Damit wissen wir zu jeder Sequenz genau die Zustandsfolge
  - Also: Maximum Likelihood Schätzung (hier ohne Pseudo-Counts)

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

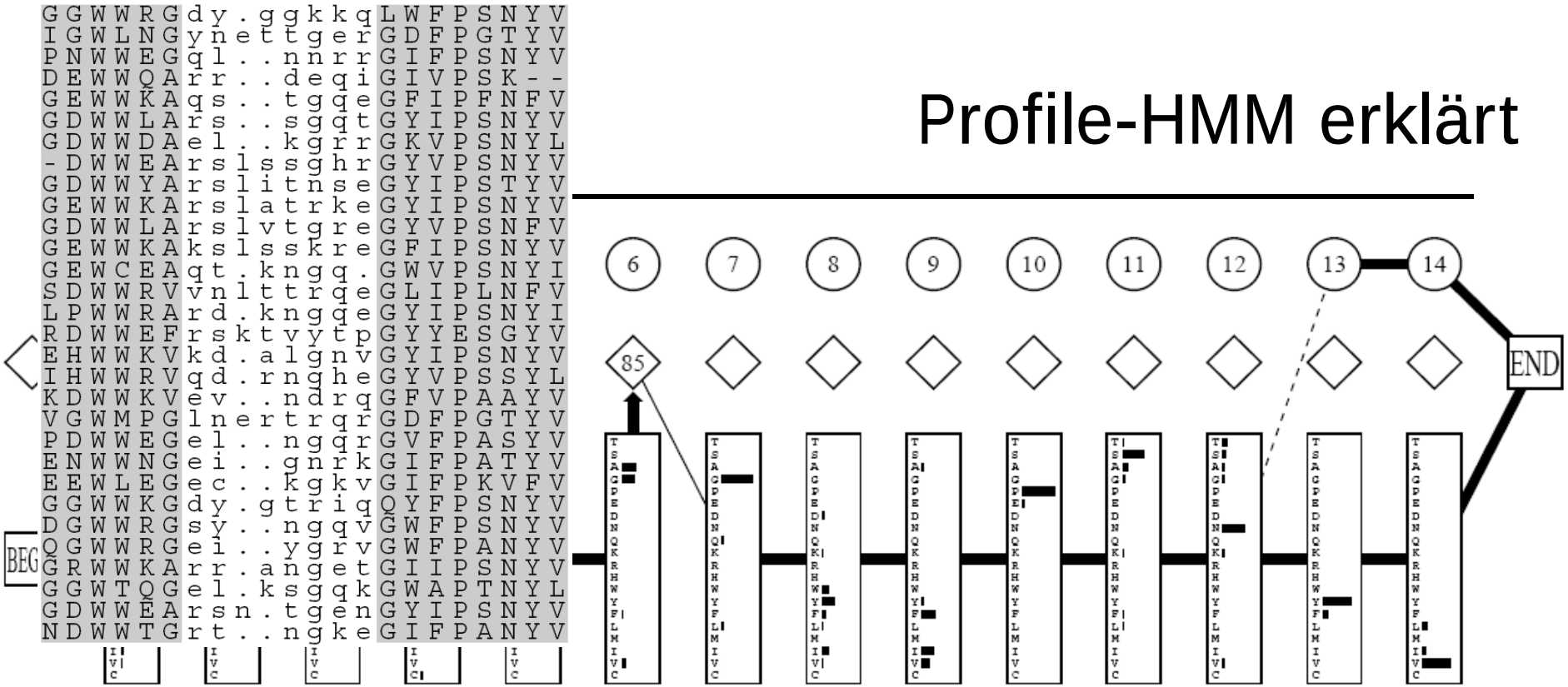
Profile-HMM erklärt





- 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



- 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?