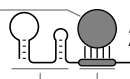
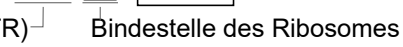
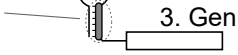
1 Über Riboswitches

Guanin-Riboswitches

Zustand	Problem	Was passiert	Ergebnis
• Zu wenig Guanin	Zelle verhungert	1. Ribosome bindet mRNA  2. Gen exprimiert  Riboswitch (in 5′ UTR) Bindestelle des Ribosomes	Mehr Guanin
• Genug Guanin	Gen-Expression Vergeudet Energie	1. Guanin bindet Riboswitch, ändert Struktur  2. Ribosom bindet nicht  3. Gen blockiert	Guanin bleibt, spart Energie

2 Definition Paar-HMM

Ein Paar-HMM besteht aus:

- Zuständen $S_1, S_2, \ldots, S_n$
  - Der Startzustand ist S_1
- Übergangswahrscheinlichkeiten $T(i,j)$ (Wkt. des Übergangs von Zustand S_i auf S_j)
- 2 Ausgabealphabete für die 2 Sequenzen. Die Alphabete können sich gleichen (z.B. beide entsprechen den 20 herkömmlichen Aminosäuren) oder abweichen (z.B. das 1. Alphabet entspricht den 20 Aminosäuren, das 2. Alphabet entspricht den 4 Nukleotiden)
  - Das Paar-HMM kann auch ein Gap in die linke bzw. rechte Sequenz ausgeben, was wir mit dem Symbol '-' schreiben.
  - Das Paar-HMM kann auch gleichzeitig in beide Sequenzen ein Gap ausgeben. Aber wir gehen davon aus, dass alle doppelten Gaps automatisch gelöscht werden, als wäre nichts ausgegeben.
- Ausgabewahrscheinlichkeiten $E(i,c,d)$ (Wkt. in Zustand S_i Symbol c in die linke Seq. und Symbol d in die rechte Seq. auszugeben) c oder d könnte auch ein Gap sein. z.B. $E(i, \mathrm{M}, -)$ entspricht der Wkt. ein 'M' in die linke Sequenz und ein Gap in die rechte Sequenz in Zustand S_i auszugeben.

Im Context des Viterbi-Algorithmus, definieren wir:

- Die Eingabesequenzen: x^L (linke Seq.) und x^R (rechte Seq.)
- Tabelle $A(i,k,l)$: Wkt. des wahrscheinlichsten Pfades, der
  - mit dem Startzustand S_1 anfängt,
  - auf den Zustand S_i endet,
  - die ersten k Symbole von x^L ausgibt (d.h. $x^L_1 x^L_2 \cdots x^L_k$), und
  - die ersten l Symbole von x^R ausgibt (d.h. $x^R_1 x^R_2 \cdots x^R_l$)

3 Viterbi-Algorithmus mit Scores/Logarithmen

Viterbi-Algorithmus für Paar-HMM mit Scores oder Logarithmen. Das heißt, wir gehen davon aus, dass die Werte in $T(i,j)$ und $E(i,c,d)$ Scores sind, die zusammen addiert werden sollten.

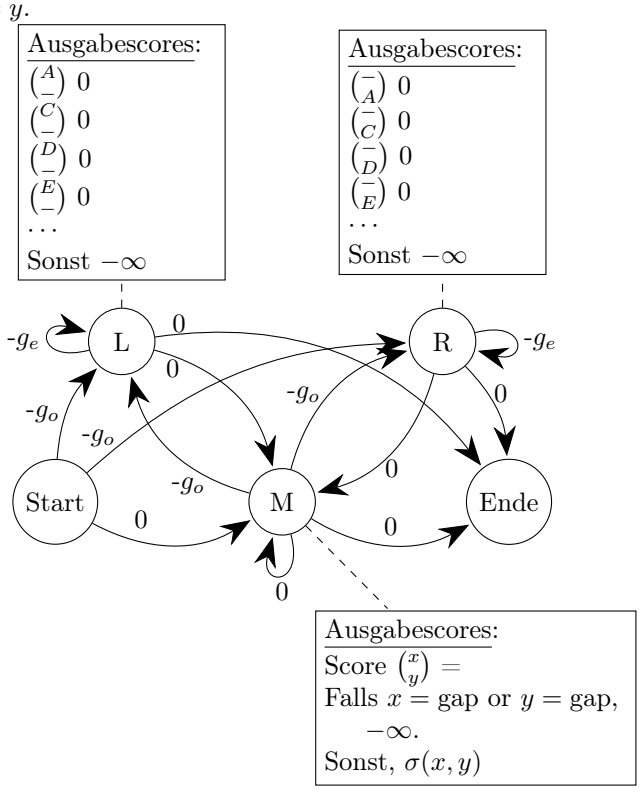
Der hier dargestellte Algorithmus kann mit dem Fall nicht umgehen, in dem sowohl c als auch d Gaps sind. Doch es ist möglich, einen solchen Algorithmus zu machen.

{Zunächst alle Zellen bekommen den Wert Null}
for all $i \in \{1,2,\ldots,n\}, k \in \{0,\ldots,|x^L|\}, l \in \{0,\ldots,|x^R|\}$ **do**
  $A(i,k,l) = -\infty$
end for
{Initialisierung vom Startzustand}
$A(1,0,0) = 0$
{Initialisierung erster Zeile und erster Spalte in A.}
for $k = 1,2,\ldots,|x^L|$ **do**
  for $i = 2,\ldots,n$ {Startzustand S_1 ist hier nicht nötig} **do**
    $A(i,k,0) = \max_{j \in \{1\ldots,n\}} A(j,k-1,0) + T(j,i) + E(i,x^L_k,-)$ {das entspricht dem normalen Viterbi-Algorithmus}
  end for
end for
for $l = 1,2,\ldots,|x^R|$ **do**
  for $i = 2,\ldots,n$ **do**
    $A(i,0,l) = \max_{j \in \{1\ldots,n\}} A(j,0,l-1) + T(j,i) + E(i,-,x^R_l)$
  end for

end for
{Endlich ergänzen wir den Rest der Tabelle A.}
for $k = 1,2,\ldots,|x^L|$ **do**
  for $l = 1,2,\ldots,|x^R|$ **do**
    for $i = 1,\ldots,n$ **do**
      Folgende Formel erinnert an Algorithmen wie Needleman-Wunsch oder Smith-Waterman. Das ist kein Zufall.
      $A(i,k,l) = \max_{j\in\{1\ldots,n\}} \begin{cases} A(j,k-1,l-1) + T(j,i) + E(i, x_k^L, y_l^R) \\ A(j,k,l-1) + T(j,i) + E(i,-,y_l^R) \\ A(j,k-1,l) + T(j,i) + E(i,x_k^L,-) \end{cases}$
    end for
  end for
end for

4 Paar-HMM: Gotoh-Algorithmus für Proteine

$\sigma(x,y)$ ist eine Score-Funktion für Aminosäuren x und y. Z.B. $\sigma(x,y) = +1$ falls $x = y$ und $\sigma(x,y) = -1$ falls $x \neq y$.



5 Genetischer Code

Diese Codone benutzen wir in der Vorlesung:

Nuk.	Aminosäure	
	Buchstabe	Name
AAA	K	Lysin
AAC	N	Asparagin
ACA	T	Threonin
ACC	T	Threonin
CAA	Q	Glutamin
CCA	P	Prolin
CCC	P	Prolin

6 Beispiel bzgl. (Bundschuh, 2004)

Organismus	Beschreibung	Sequence
Schleimpilz-2	Proteinseq. *nad* 2	KTQ *(In Eingabe)*
Schleimpilz-2	Proteinseq. *nad* 2 mit Gaps	`--K--T--Q`
P. polycephalum	Proteinsequenz von vorhergesagter mRNA-Sequenz	`--K--T--P`
P. polycephalum	mRNA-Sequenz mit vorhergesagten C-Insertionen	`AAAACACCAAA`
P. polycephalum	DNA-Sequenz vom Genom mit Gaps	`AAAA-A--AAA`
P. polycephalum	DNA-Sequenz vom Genom	AAAAAAAA *(In Eingabe)*

Wir alignen eine Nukleotidsequenz mit einer Proteinsequenz, und müssen berücksichtigen: (1) den genetischen Code, (2) C-Insertionen und (3) normale Gaps wie im Gotoh-Algorithmus.