

Lokale Sequenzähnlichkeit

```
LHEVEDRYVVLVDAELTFVRQLEIYRASRDKGLRVLVFLIYGGSTEEQRVLTALRKEEA
L E+ P Y+++++ +++F+RQ+E+Y+A      +VVF+ YG S EEQ +LTA+++EK+A
LQEMMPSYIIMFEPDISFIRQIEVYKAIKVDLQPKVYFMYGESIEEQLTAIKREKDA

FEKLIKASIMVVFEBREGDRETNLDLVRGTASAD-----VSTDFKAGGQE--QNG
F KLERE A++      ETN DL      A+      +TR AGGQ+  N
FTKLIKENANL-----SHFPTNEDLSHYKNLAERKLLKLRKSNTRNAGGQQGFHNL

TQQSIVVMREFRSELP SLIHRGIDIEPVTLEVDYILTDMCVKRSISDLIGSLNNG
TQ  +VVD REF + LP L+R GI + P L VGDY+TR+C+ERKSIDLIGSL N
TQDVVIVDTREFNAPLGLLYRGIKRVIPCHLTVGDYVITPDICLERKSIDLIGSLNN 4
```

Genomische Datenanalyse

9. Kapitel

Globale Sequenzähnlichkeit

```

M C D V E E G E K I I F I R E C S
Cytochrome C Human: ATG GGT GAT GTT GAG AAA GGC AAG AAG ATT TTT ATT ATG AAG TGT TGC
101 110 120 130 140 150 160 170 180 190 200 210 220 230 240
Cytochrome C Mouse: ATG GGT GAT GTT GAG AAA GGC AAG AAG ATT TTT ATT ATG AAG TGT TGC
25 35 45 55 65 75 85 95 105 115 125 135 145 155

S C R T V F E E K G G E R E T G P S
Cytochrome C Human: CAG TGC CAG GGC GTT GAA AAG GGA GGC AAG CAG AAG ACT GGC CGA AAT
161 170 180 190 200 210 220 230 240 250 260 270 280 290 300
Cytochrome C Mouse: CAG TGC CAG GGC GTT GAA AAG GGA GGC AAG CAG AAG ACT GGC CGA AAT
31 41 51 61 71 81 91 101 111 121 131 141 151 161 171

L N G L F G R E R T E G A P G T S
Cytochrome C Human: GTC CAG GAT GTC TTT GGG GAG AAG ACA GAT GAG GGC GCT GCA TGC TCT
301 310 320 330 340 350 360 370 380 390 400 410 420 430 440
Cytochrome C Mouse: GTC CAG GAT GTC TTT GGG GAG AAG ACA GAT GAG GGC GCT GCA TGC TCT
45 55 65 75 85 95 105 115 125 135 145 155 165 175

T T A A R E E R K G I V G E D T
Cytochrome C Human: TAC ACA GGC GGC AAT AAG AAC AAA GGC ATC ATC TAC TAC GAG GAT ACA
441 450 460 470 480 490 500 510 520 530 540 550 560 570 580
Cytochrome C Mouse: TAC ACA GGC GGC AAT AAG AAC AAA GGC ATC ATC TAC TAC GAG GAT ACA
51 61 71 81 91 101 111 121 131 141 151 161 171 181

L R E T L E R P R E T I P G T R
Cytochrome C Human: GTC ATC GAG TAT TTA GAG AAT GTC AAA AAG TAC ATC GAT GCA ACA AAA
591 600 610 620 630 640 650 660 670 680 690 700 710 720 730
Cytochrome C Mouse: GTC ATC GAG TAT TTA GAG AAT GTC AAA AAG TAC ATC GAT GCA ACA AAA
61 71 81 91 101 111 121 131 141 151 161 171 181 191

N I F Y G I E K K E E N A D L I
Cytochrome C Human: ATC ATC TTT GTC GGC ATT AAG AAC AAG GAA AAG GCA GAC GTA ATA
741 750 760 770 780 790 800 810 820 830 840 850 860 870 880
Cytochrome C Mouse: ATC ATC TTT GTC GGC ATT AAG AAC AAG GAA AAG GCA GAC GTA ATA
81 91 101 111 121 131 141 151 161 171 181 191 201 211

A T L E E K A T R E
Cytochrome C Human: GCT TAT CTC AAA AAA GCT ACT AAT GAG
891 900 910 920 930 940 950 960 970 980 990 1000
Cytochrome C Mouse: GCT TAT CTC AAA AAA GCT ACT AAT GAG
91 101 111 121 131 141 151 161 171 181 191 201 211

```

Zwei Cytochrome C Sequenzen: Eine vom Menschen und eine aus der Maus.

Die Sequenzen sind gleich lang, man kann sie von vorne bis hinten untereinander schreiben, und beobachtet **nur wenige Mismatchpositionen** in der DNA und noch weniger in der Proteinsequenz.

Warum können das weniger sein?

Redundanz des genetischen Codes. Verschiedene Triplets für die gleiche Aminosäure.

Evolutionäres Spektrum

Species	Number of mismatches compared to the human sequence
Chimpanzee	0
Mouse	9
Fruit fly	29
Wheat	43
Yeast	51

Nimmt man die Cytochrome C Sequenzen von **weiter entfernten Eukaryonten**, häufen sich mehr **Mismatchpositionen** an.

Die globale Sequenzähnlichkeit ist aber immer noch klar erkennbar.

Entfernt verwandte Sequenzen

Wie sieht es bei Prokaryonten aus? Sie besitzen auch ein Cytochrome C.

```

Cytochrome C human: GDVEKGGkklifinkesqchtvekggkhtgmalhglfgrkTQQAPG
YSYTAANKHKGILWGEDTLMEYLEMPPKTYLPGTKMIFVGIKGGKKE
RADLIAYLKGATNE

Cytochrome C550
Bacillus subtilis: MKGNPIIPFLLIIVLGIIGLITFFLSVIGLDDSRRIASGGESKSAEK
KDNANASPeeykanciachgenyevagpslkgvdkkkWAETKT
KTEKGGHGMPSGLVPADKLDMAEMVSKIK

```

```

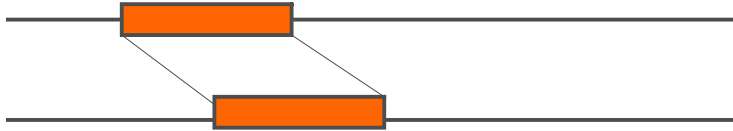
10      20      30
Cytochrome C human: ... KKIFIMKCSQCHTVEKGGKHTGPNLHGLFGRK ...
... ..i i i . i ..i.i.i. i
Cytochrome C550 Bacillus subtilis: ... EEIYKANCIAICHGENYDQ--VSGPSLKGWGDNK ...
60      70      80

```

Es gibt nur noch relativ **kurze Segmente**, in denen die gemeinsame Abstammung erkennbar ist. Und da auch nur noch **sehr schwach**.

Lokale Sequenzähnlichkeit

Bei entfernt verwandten Sequenzen beobachten wir nur noch lokale Sequenzähnlichkeiten.



Proteinähnlichkeit

Wann sind Proteinsegmente ähnlich?

Wenn sie viele identische Aminosäuren aufweisen ...

... aber in dem Beispiel sind gar nicht so viele Aminosäuren identisch. Wie wurden die Segmente dann identifiziert?

Es wurde beobachtet, daß im Laufe der Evolution **bestimmte Aminosäuren häufig gegen andere ausgetauscht werden**.

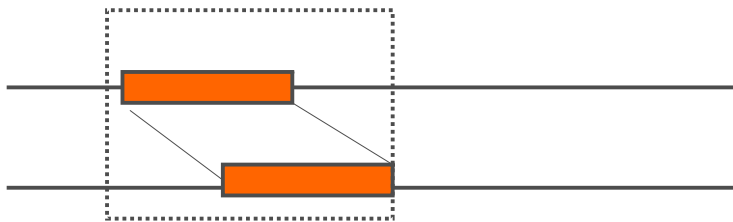
Diese Austausche lassen die Proteine funktionsfähig. Sie sind kompatibel mit der Funktion und der räumlichen Struktur der Proteine.

Besonders häufig sind Austausche von Aminosäuren, die auch **chemisch ähnliche Eigenschaften** haben: hydrophobe durch hydrophobe, große durch große, polare durch polare, etc.

Austausche von chemisch unterschiedlichen Aminosäuren sind dagegen selten.

Ähnlichkeit-Score

Angenommen wir könnten die chemische Ähnlichkeit von Aminosäuren in einer vernünftigen Form quantifizieren ... sagen wir mit Hilfe eines **Ähnlichkeitsscores** $S(x,y)$, der je zwei Aminosäuren x und y eine Zahl zuordnet. Hohe Zahlen für ähnliche Paare und niedrige für unähnliche Paare.



Wie könnte man dann die lokale Ähnlichkeit zweier Sequenzen bemessen?

... KLYMCWA ...

$S(K,K)+S(L,A)+\dots+S(A,L)$

... KAWMCYL ...

Was ist $S(x,x)$?

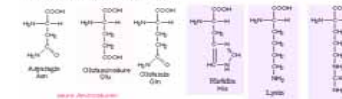
Ähnlichkeit von Aminosäuren

Wie mißt man die Ähnlichkeit von Aminosäuren?

Aminosäuren mit hydrophoben Resten



Aminosäuren mit hydrophilen Resten



Basische Aminosäuren



Neutrale Aminosäuren



- hydrophob A, L, I, V, M, C, W, F, P
- hydrophil D, E, K, R, H
- groß W, R
- klein G, A, S
- sauer D, E
- basisch K, R, H
- aromatisch W, F, Y
- polar S, T, Y, N, Q

Ein Biochemisches Problem ...

In einem konservierten Bereich einer Proteinsequenz erwartet man entweder identische Aminosäuren, oder aber alignierte Aminosäuren, die chemisch ähnliche Eigenschaften haben.

Wie quantifiziert man die chemische Ähnlichkeit von Aminosäuren?

Wie verrechnet man ihre Größe mit ihrem PH-Wert und ihrer Hydrophibizität?

Und wie würde man eine chemische Ähnlichkeitsskala mit der Austauschbarkeit der Aminosäuren in Bezug setzen?

... mit einer evolutionsbiologischen Lösung

Trick:

Man geht umgekehrt vor: Man definiert die Ähnlichkeit von Aminosäuren aufgrund ihrer Austauschbarkeit.

Hat man einen großen Datensatz teilweise konservierter Sequenzen, kann man beobachten, welche Austausche häufig sind und welche eher selten ... dies gibt dann ein Ähnlichkeitsmaß.

Ähnlichkeitsmatrix

Ziel ist eine Ähnlichkeitsmatrix:

20 x 20

für jedes Paar von Aminosäuren ein Ähnlichkeitswert.

Es gibt mehrere Ansätze :

- PAM (Magaret und Dayhoff 1978)
- BLOSUM (Henikoff und Henikoff 1993)
- VT (Müller und Vingron 2000)

BLOSUM

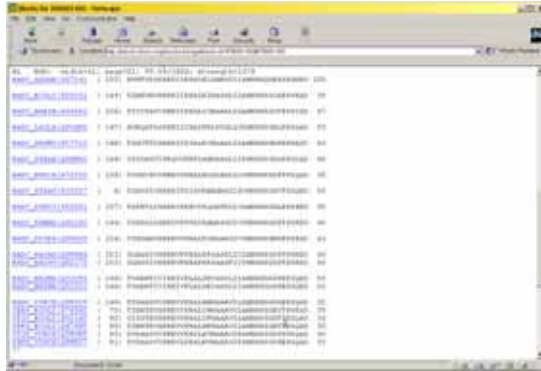
Der BLOSUM Matrix beruht auf der BLOCKS-Datenbank:

<http://www.blocks.fhcrc.org/>



BLOCKS

BLOCKS enthält **multiple Alignments von konservierten Segmenten**, die man in mehreren Proteinen finden kann.



Insertionen und Deletionen dürfen in BLOCKS nicht auftreten.

Von BLOCKS zu BLOSUM

Das sind genau die Daten, die man braucht, um die Ähnlichkeit von Aminosäuren zu charakterisieren.

Man beobachtet **Austausche** (Substitutionen) von Aminosäuren in den Blocks.

Es gibt häufig beobachtbare Substitutionen und seltene.

Wir müssen sie nur **auszählen**:

Dabei gibt es aber noch einiges zu beachten:

Symmetrie ?

Wir wollen jedem Paar von Aminosäuren (x,y) eine Zahl zuordnen, die beschreibt wie „*einfach*“ ein x durch ein y ersetzt werden kann ... es gibt 20×20 Paare.

Macht es Sinn, dem Paar (L,S) einen anderen Wert als dem Paar (S,L) zuzuordnen? Kann es z.B. sein, daß ein L eher zu einem S werden kann als ein S zu einem L?

Mutation vs. Substitution !

Nein!

Es gibt einen Unterschied zwischen einer **Mutation** und einer **Substitution**. Eine Mutation ist gerichtet. Man hat einen Vorgänger und einen Nachfolger:

z.B:

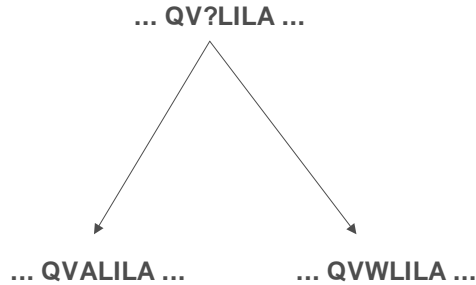
Vorgänger: ... KALMY ...

Nachfolger: ... KALPY ...

Hier ist ein M in ein P mutiert und nicht umgekehrt.

Mutation vs. Substitution !

In den BLOCKS beobachten wir Sequenzen, die einen gemeinsamen Vorfahren haben. Aber nicht diesen Vorfahren.



Wir beobachten einen Aminosäureaustausch $A \leftrightarrow W$ in Sequenzen, die „Zeitgenossen“ sind.

Dem kann eine Mutation $A \rightarrow W$ oder eine Mutation $W \rightarrow A$ zugrunde liegen. Welche, wissen wir nicht.

Symmetrie !

Es macht also keinen Sinn, zwischen dem Paar (x,y) und dem Paar (y,x) zu unterscheiden.

Es kann sein, daß der Mutationsprozess unsymmetrisch ist, wir haben aber keinen Zugang zu Daten die das belegen würden. Deshalb können wir nur einen symmetrischen Ähnlichkeitsscore aus Sequenzdaten ableiten

Außerdem würde ein unsymmetrischer Score auch der Idee von Aminosäureähnlichkeiten nicht entsprechen.

Selbstähnlichkeit

Macht es Sinn, einem Paar (x,x) einen Wert zuzuordnen?

Ja!

Es gibt Aminosäuren, die häufiger mutieren als andere. Weil die Menge von 20 Aminosäuren für eine Aminosäure mehr chemisch ähnliche Alternativen bereitstellt als für andere.

z.B. nur Cystein kann Cysteinbrücken bilden.

Und in der Tat findet man relativ selten, daß ein Cystein gegen eine andere Aminosäure ausgetauscht wurde.

Wir brauchen also 210 Werte, die Ähnlichkeit oder Austauschbarkeit beschreiben.

Ähnlichkeit vs. absolute Substitutionshäufigkeit

Betrachten wir zwei Paare von Aminosäuren

1. (L, A)
2. (W, Y)

Man beobachtet das Paar (L,A) weitaus öfters in den Blocks als das Paar (W, Y).

Folgt daraus schon, daß (L,A) sich ähnlicher sind als (W, Y)?

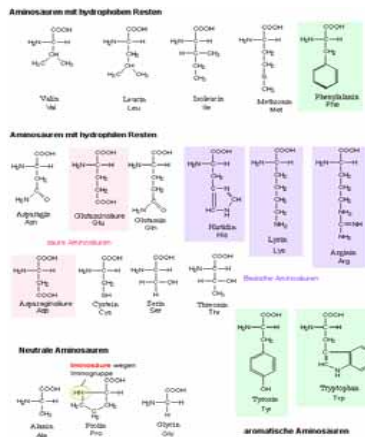
L: Leucin

A: Alanin

W: Tryptophan

Y: Tyrosin

Ähnlichkeit vs. absolute Substitutionshäufigkeit



AA	rel H'keit	% L H'keit
L	0.098	100
A	0.077	78.9
S	0.071	72.4
V	0.067	68.5
G	0.066	67.1
E	0.063	64.9
K	0.059	60.6
I	0.059	60.6
T	0.057	58.1
D	0.054	55.1
R	0.05	51.3
P	0.048	48.8
N	0.046	47.4
F	0.041	42.3
Q	0.041	41.9
Y	0.032	33.2
M	0.022	22.6
H	0.022	22.4
C	0.014	14.8
W	0.013	13

Problem: Aminosäurehäufigkeiten

Beide Paare von Aminosäuren sind sich relativ ähnlich.

Daß (L,A) häufiger zu beobachten ist als (Y,W) liegt wohl eher daran, daß **Y und W seltene Aminosäuren** sind, was man von **L und A nicht behaupten kann**.

Was ist zu tun?

Hintergrundmodell

Wie wahrscheinlich ist es, an einer festen Position das Paar (x,y) zu beobachten, wenn man einfach nur zwei nicht verwandte Proteine untereinander schreibt?

$p_x p_y$ wobei p_x und p_y die relativen Häufigkeiten der Aminosäuren x und y sind.

Wie definiert man jetzt einen Ähnlichkeitsscore?

Vorschlag:

$$S(x, y) = \log \left(\frac{p_{xy}}{p_x p_y} \right)$$

wobei p_{xy} die relative Häufigkeit des Paares (x,y) in der Blocksdatenbank ist.

Von der Ähnlichkeit von Aminosäuren zurück zur Ähnlichkeit von Proteinsegmenten

Was machen wir eigentlich, wenn wir lokale Sequenzähnlichkeit mit diesem Score bewerten?

Wir haben ein lokales Alignment (ohne gaps):

...x₁ x₂ ... x_n ...

...Y₁ Y₂ ... Y_n ...

$$S(\text{Alignment}) = S(x_1, y_1) + \dots + S(x_n, y_n)$$

$$= \sum \log \left(\frac{p_{x_i y_i}}{p_{x_i} p_{y_i}} \right) = \log \left(\frac{\prod p_{x_i y_i}}{\prod p_{x_i} p_{y_i}} \right)$$

Ähnlichkeit als Log-Likelihood-Ratios

Modell für verwandte Sequenzen:

i.i.d. Folge von Paaren von Aminosäuren:

Randverteilung der X_i : $P[X_i=(x,y)] = p_{xy}$

Modell für nicht verwandte Sequenzen:

i.i.d. Folge Y_i : $P[Y_i=(x,y)] = p_x p_y$

$$\log \left(\frac{\prod p_{x_i y_i}}{\prod p_{x_i} p_{y_i}} \right)$$

Likelihood des lokalen Alignments im Modell verwandter Sequenzen

Likelihood des lokalen Alignments im Modell nicht verwandter Sequenzen

Nahe vs. ferne Verwandtschaft

Der Score ist ein **Likelihood Ratio** zwischen einem **Modell für das was wir suchen** (verwandte Sequenzabschnitte) und einem **Hintergrundmodell**.

Also völlig analog zu dem Score bei den Splicestellen.

Problem: Was genau modellieren wir, wenn wir verwandte Sequenzen modellieren?

Wie sieht p_{xy} aus, wenn man nah verwandte Sequenzen in den Blocks betrachtet ?

Viele Identitäten (p_{xx} groß, p_{xy} klein)

Wie sieht es aus, wenn man entfernt verwandte Sequenzen betrachtet ?

p_{xx} immer noch größer als p_{xy} , aber weniger Unterschied.

Verwandtschaftsgrad

Man kann nur einen Verwandtschaftsgrad auf einmal modellieren.

Den Verwandtschaftsgrad kann man z.B. durch den **Prozentsatz identischer Positionen** im Alignment charakterisieren.

Wir setzen für die p_{xy} relative Häufigkeiten beobachteter Paare in den Blocks-Alignments.

Für die **BLOSUM60** Scorematrix wurden Alignments mit ca. **60%** identischen Positionen dafür ausgezählt.

Bei **BLOSUM80** waren es **80%** Identität etc.

Verwandtschaftsgrad

Beim Scoring mit der **BLOSUM60** Matrix vergleichen wir die Likelihood zweier Modelle

1. Die Sequenzen haben den Verwandtschaftsgrad 60% Identität
2. Die Sequenzen sind nicht verwandt.

Was passiert mit Sequenzen, die 50% Identität haben?

Die sollten eher ins erste Modell passen.

Alignmentscore

$$BLOSUM60(x, y) = \log \left(\frac{p_{xy}(60)}{p_x p_y} \right)$$

Dieser Score macht auch für die „Selbstähnlichkeit“ $S(x,x)$ oder besser gesagt für die generelle Austauschbarkeit einer einzelnen Aminosäure Sinn.

Damit wissen wir, wie wir einem Paar alignierter Sequenzabschnitte (ohne Gaps) einen Ähnlichkeitsscore zuordnen:

... x_1 x_2 ... x_n ...

... Y_1 Y_2 ... Y_n ...

$$\text{Score(Alignment)} = \sum BLOSUM60(x_i, y_i)$$

BLOSUM60

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	1	4						
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	2	2	4					
V	-1	-2	0	-2	0	-3	-3	-3	-2	-3	-3	-2	1	3	1	4				
F	-2	-2	-4	-2	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6				
Y	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	3	7			
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-3	-3	-1	-3	-2	-3	1	2	11	

Die Einträge wurden auf ganze Zahlen gerundet. Dies geschieht, um den hohen Rechenaufwand bei Datenbanksuchen zu reduzieren.

Positive Einträge:
Das Paar (x,y) sieht man in 60% - Identität-Blocks häufiger als man durch Zufall erwartet.

Negative Einträge:
Diese Paare sieht man in den Blocks seltener als erwartet.

Diagonal Einträge:
positive Werte, generelle Austauschbarkeit der Aminosäure.

Lokales Alignment

Im Allgemeinen weiß man aber nicht, wo konservierte Segmente sind. Die will man ja gerade suchen.

Was muß man dazu tun?

Optimales lokales Alignment (ohne gaps) zweier Sequenzen:

Finde das Paar von Segmenten (jeweils eines in jeder Sequenz), daß den höchsten Score besitzt. Dieser Score wird als Ähnlichkeitsmaß der beiden Sequenzen verwendet.

(Lokales Ähnlichkeitsmaß)

Beachte:

Zunächst wird über alle möglichen Segmentpaare **optimiert**, und dann wird der Score für das optimale Segmentpaar verwendet.

% Identität

Wozu braucht man dieses neue Maß für lokale Sequenzähnlichkeit? Warum nimmt man nicht wieder %Identität als Ähnlichkeitsmaß, wie es bei der Auswahl des Datensatzes für die BLOSUM 60 Matrix getan wurde?

Der Prozentsatz Identität ist bei kürzeren lokalen Alignments höher als bei langen.

Als Extremfall kann man sich überlegen, nur einen Buchstaben der ersten Sequenz mit dem gleichen Buchstaben in der zweiten Sequenz zu alignieren. Also Segmente der Länge 1. Das sind dann 100% Identität. Man ist aber an längeren konservierten Bereichen interessiert.

% Identität vs. Alignmentscore

Der optimale lokale Alignmentscore steigt, solange man das lokale Alignment um Bereiche verlängert, für die die Likelihood im Modell für verwandte Sequenzen höher ist als im Hintergrundmodell, auch wenn der Prozentsatz Identität dadurch fällt.

Als Maß für Sequenzkonservierung bei gegebenem Alignment ist der Prozentsatz Identität brauchbar (auch wenn es bessere Alternativen gibt). Zum Suchen konservierter Bereiche ist er im Gegensatz zum Score unbrauchbar.

Large scale Effekte

Der optimale lokale Alignmentscore ist immer positiv (obwohl die Scorematrix auch negative Einträge enthält)

Man braucht ja wieder nur zwei identische Aminosäuren zu matchen. Dieses lokale Alignment der Länge 1 hat dann einen positiven Score

Im allgemeinen sind die lokalen Alignments aber viel länger und haben höhere Scores, auch bei nicht verwandten Sequenzen.

Das bedeutet, daß man immer Segmente finden kann, die eher nach Verwandtschaft als nach Zufall aussehen.

Es gibt viele mögliche Segmente, die man miteinander kombinieren kann.

Deswegen findet man auch unter nicht verwandten Sequenzen ähnliche Segmente. Diese gehen aber nicht auf evolutionäre Konservierung zurück, sondern sind Zufall (**Large scale Effekte**).

Zufällige Sequenzähnlichkeit

Wann ist eine Sequenzähnlichkeit ein Indiz für die Homologie der Sequenzen ?

Wann ist eine Sequenzähnlichkeit signifikant ?

Wie hoch kann der Score durch Zufall werden?

Wie geht man dieses Problem an?

Wir brauchen ein Modell für **zufällige Sequenzähnlichkeit**.

Lokales Alignment **zufälliger** Sequenzen

Nehmen wir zufällige Sequenzen und alignieren sie, so haben sie ähnliche Segmente. Das ist auf alle Fälle Zufall und keine Spur der Evolution.

Sequenz X: X_1, \dots, X_n i.i.d. mit $X_i \sim (p_1, \dots, p_{20})$

Sequenz Y: Y_1, \dots, Y_m i.i.d. mit $Y_i \sim (p_1, \dots, p_{20})$

Beide Sequenzen werden unabhängig voneinander generiert.

Zufälliger optimaler lokaler Alignment Score:

$$H = \max S(X', Y'),$$

wobei X' und Y' Segmente von X und Y sind.

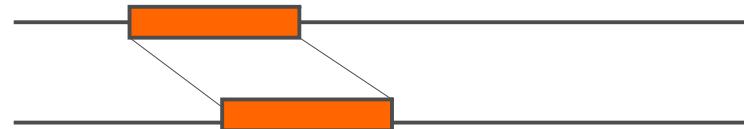
Konservierte Bereiche vs. ...

In Kapitel 7 hatten wir i.i.d. Sequenzen generiert und untereinander geschrieben. Dann hatten wir überlegt, ob es durch Zufall Bereiche gibt, in denen die obere Sequenz identisch mit der unteren ist.



... vs. lokales Alignment

Hier berechnen wir von den Zufallssequenzen zunächst das optimale lokale Alignment, daß heißt wir vergleichen nicht übereinander liegende Segmente sondern beliebige Segmente (größerer Suchraum).



Zufälliger Alignmentsscore

H ist eine Zufallsvariable:

H ist deterministisch abhängig von den X_i und Y_i .

Wie ist H verteilt?

$P[H < t] = ?$

$$= 1 - P[H \geq t]$$

Falls $H \geq t$ gilt, gibt es ein lokales Alignment mit $\text{Score} \geq t$.

Zu diesem Alignment gehören zwei Sequenzsegmente s_1, s_2

Diese Segmente haben Startpunkte n_1, n_2 in den jeweiligen Sequenzen.

Es gibt $n \times m$ mögliche Startpunkte für ein lokales Alignment mit $\text{Score} \geq t$ in den Sequenzen:

X_1, \dots, X_n und Y_1, \dots, Y_m .

Spezialfall: Das längste gemeinsame Wort

Betrachten wir vorübergehend einen etwas einfacheren Extremfall:

Statt der BLOSUM60 Matrix betrachten wir eine Scorematrix S , für die $S(x,x) = 1$ und $S(x,y) = -\infty$ für $x \neq y$ gilt.

Was ist das optimale lokale Alignment für diese Scorematrix?

Das **längste gemeinsame Teilwort** beider Sequenzen!

$H \geq t$ bedeutet, daß die Sequenzen ein Wort der Länge t gemeinsam haben.

Das längste gemeinsame Wort

Wie wahrscheinlich ist es, daß beide Sequenzen mit einem gemeinsamen Wort der Länge t beginnen?

Das ist gleichbedeutend damit, daß eine Bernoullisequenz mit einem Headrun der Länge t startet.

Das geschieht mit Wahrscheinlichkeit p^t .

Wobei p die Wahrscheinlichkeit $P[X_i=Y_i]$ ist.

$$P[X_i=Y_i] = \sum_{i=1}^{20} p_{ii}$$

Poissonapproximation & Declumping

Sei W die Anzahl gemeinsamer Wörter mit Mindestlänge t :

Welche Verteilung hat W ?

Lange gemeinsame Wörter sind **seltene Ereignisse**:

Also für t groß genug könnte W **poissonverteilt** sein.

Gibt es ein **Declumping Problem**?

Auf alle Fälle: Hat man ein gemeinsames Wort der Länge $t+\Delta$, wächst W direkt um Δ .

Das ist die gleiche Situation wie bei den Headruns. Es empfiehlt sich, Declumping anzuwenden: Wir zählen nur gemeinsame Wörter, die sich nicht verlängern lassen.

Intensität

Welche Intensität hat die declumpte Zufallsvariable W ?

$$\lambda = E[W] = p^t \{ (n-t)(m-t)(1-p) + 1 \}$$

mögliche Startpunkte
Randeffekte Declumping:
 $X_{n-1} \neq Y_{m-1}$

Für eine poissonverteilte Zufallsvariable W gilt $E[W]=\lambda$.
 $E[W] =$
 Erfolgswahrscheinlichkeit \times
 Gesamtanzahl Versuche

Von der Verteilung von W zur Verteilung von H

Haben Sequenzen kein Wort der Länge t gemeinsam, dann ist $W=0$

$$P[H \geq t] = 1 - P[W=0]$$

W ist poissonverteilt mit Intensität λ .

$$P[W = k] = \frac{e^{-\lambda} \lambda^k}{k!}$$

$$\text{Also: } P[W=0] = e^{-\lambda}$$

$$P[H \geq t] \approx 1 - e^{-\lambda}$$

Einsetzen und umformen

$$\begin{aligned}
 P[H \geq t] &\approx 1 - e^{-\lambda} \\
 &= 1 - \exp\left(-p^t(\{(n-t)(m-t)(1-p) + 1\})\right) \\
 &\approx 1 - \exp\left(-\gamma n m e^{-\frac{t}{\theta}}\right)
 \end{aligned}$$

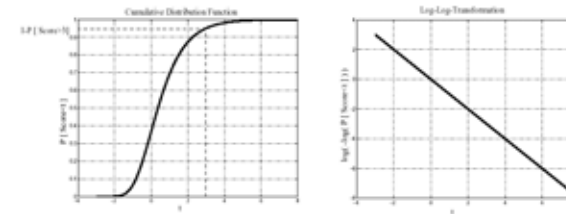
Wobei γ und θ von p abhängen, die Randeffekte und das Declumping auffangen, aber nicht von den Sequenzlängen n und m abhängen.

Was ist das für eine Formel?

Standard Extremwertverteilung

Eine kontinuierliche Zufallsvariable G mit der Verteilungsfunktion ... $P[G < t] = e^{e^{-t}}$

... heißt **standard-extremwertverteilt**.



Trägt man die y-Achse bei der Verteilungsfunktion von G doppelt logarithmisch (loglog) auf, ergibt sich eine Gerade.

Familie der Extremwertverteilungen

Wie die Normalverteilungen bilden die Extremwertverteilungen eine parametrisierte Familie:

Ist G standard extremwertverteilt, dann gilt für $X = \theta G + \xi$:

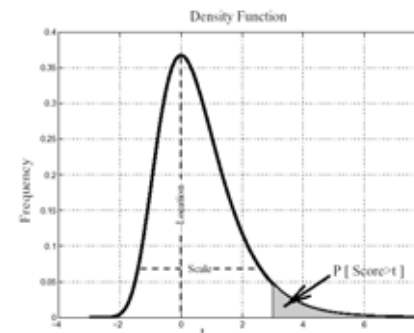
$$\begin{aligned}
 P[X < t] &= P\left[G < \frac{t - \xi}{\theta}\right] \\
 &= \exp\left(-e^{-\frac{t - \xi}{\theta}}\right) \\
 &= \exp\left(-e^{\frac{\xi}{\theta}} e^{-\frac{t}{\theta}}\right)
 \end{aligned}$$

Man sagt X ist extremwertverteilt mit Skalenparameter θ und Lageparameter ξ , und schreibt kurz $X \sim G(\xi, \theta)$.

Dichte einer Extremwertverteilung

Die dazugehörige Dichte lautet:

$$\phi(t) = \theta^{-1} e^{-(t-\xi)/\theta} \exp\left(-e^{-\frac{t-\xi}{\theta}}\right)$$



Die Verteilung ist **nicht symmetrisch**.

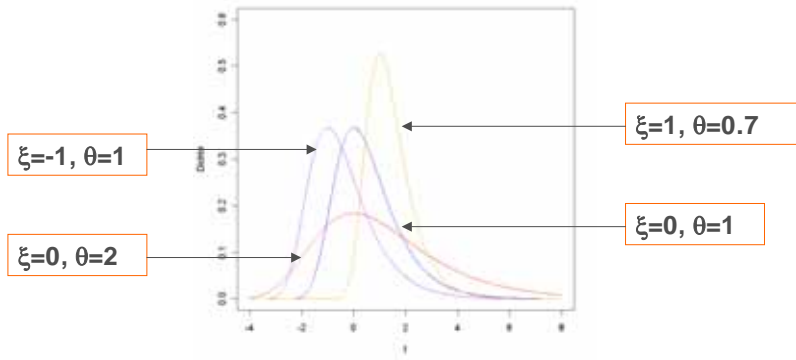
Relativ langsames Abfallen für große Werte mit Rate e^{-t}

... im Vergleich zu e^{-t^2} bei der Normalverteilung.

Abweichungen von mehreren Standardabweichungen vom Erwartungswert sind **nicht selten**.

Familie der Extremwertverteilungen

Einige Mitglieder der **parametrisierten Familie** der Extremwertverteilungen:



Erwartungswert und Varianz

Ist $X \sim G(\xi, \theta)$, dann gilt:

$$E[X] = \xi + c \theta,$$

wobei c die Eulerkonstante $c \approx 0.577$ ist.

und

$$\text{Var}(X) = 1/6 \pi^2 \theta^2.$$

Zurück zu den gemeinsamen Wörtern

Mit Poissonapproximation haben wir die Verteilung der Länge der längsten Wörter bestimmen können als:

$$P[H \geq t] \approx 1 - \exp\left(-\gamma n m e^{-\frac{t}{\theta}}\right)$$

Auch das ist eine Extremwertverteilung mit
Skalenparameter: θ

Lageparameter: $\theta \log(\gamma m n)$

Zurück zu den gemeinsamen Wörtern

Die Länge des längsten gemeinsamen Wortes ist natürlich ganzzahlig und nicht kontinuierlich, aber für die ganzzahligen Werte t , die sie annehmen kann, ist die **Approximation** mit einer Extremwertverteilung treffend.

γ und θ lassen sich aus der Verteilung der X_i und Y_i ausrechnen.

Zufällige lokale Alignmentscores

Zurück von der einfachen Scorematrix S zur BLOSUM60 Matrix ...
... oder zurück von gemeinsamen Worten zu ähnlichen
Segmenten.

Man kann zeigen, daß auch der optimale lokale Alignmentscore
extremwertverteilt ist, und daß auch hier die gleiche Formel gilt,
allerdings mit anderen Werten für die Konstanten γ und θ .

$$P[H \geq t] \approx 1 - \exp\left(-\gamma n m e^{-\frac{t}{\theta}}\right)$$

Zufällige lokale Alignmentscores

$$P[H \geq t] \approx 1 - \exp\left(-\gamma n m e^{-\frac{t}{\theta}}\right)$$

Diese Formel gibt die Wahrscheinlichkeit an, daß man eine
lokale Sequenzähnlichkeit mit Score $\geq t$ auch in zwei zufälligen
Sequenzen finden kann.

Die zufälligen Sequenzen sind ein Modell für nicht verwandte
Sequenzen. **Die Formel gibt damit auch an, wieviel „Phantom-
Ähnlichkeit“ man typischerweise in zwei nicht verwandten
Sequenzen findet.**

Die Formel spielt eine entscheidende Rolle in Datenbank-
Suchprogrammen. Mehr dazu im nächsten Kapitel.

Zusammenfassung:

Es gab nur einen neuen statistischen Begriff:
Extremwertverteilung