

Aufgabe 5 - Substitutionsmatrix

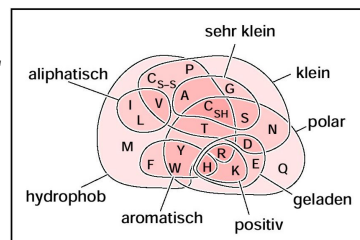
Silke Trißl

Wissensmanagement in der Bioinformatik



Substitutionsmatrizen

- Ersetzung einer Base/Aminosäure durch eine andere hat **unterschiedliche Bedeutung**
 - Basen: Auswirkungen auf kodierte Protein nicht gleichverteilt über die drei Codon-Positionen
 - Aminosäuren
 - Ersetzung mit „sehr ähnlichen“ Aminosäuren ändert Proteinstruktur kaum
 - Ersetzung mit „wenig ähnlichen“ Aminosäuren kann Struktur vollkommen ändern



PAM

- Point Accepted Mutation
 - Dayhoff et al., 1978
- Zwei Bedeutungen
 - 1 PAM – Einheit für den Abstand von Proteinsequenzen
 - PAM-X Matrix – Beobachtete Substitutionsmatrix für zwei Sequenzen die X PAM entfernt sind
- Gebräuchlichste
 - PAM250

PAM Matrizen

- Seien $(S_{1,1}, S_{2,1}) \dots (S_{1,n}, S_{2,n})$ Paare von Sequenzen, die jeweils x PAM entfernt sind. Dann berechnet sich die PAMx Matrix M_x wie folgt
 - Messe absolute Häufigkeit $f(A_i)$ für alle Aminosäuren A_i in allen Sequenzen
 - Aligniere alle Paare entsprechend der evolutionären Wahrheit
 - $S_{k,l}$ sei $S_{k,l}$ mit den durch das Alignment eingefügten Leerzeichen
 - Messe Übergangshäufigkeit $f(i,j)$ zwischen allen Paaren (A_i, A_j) , normiert auf Gesamtzahl aller Paare
 - Anzahl von Positionen k mit $S_{1,z}[k]=A_i$ und $S_{2,z}[k]=A_j$ über alle Positionen k in allen Paaren
 - Paare $(A_x, _)$ werden ignoriert
 - Übergang ist „richtungslos“; $f(i,j) = f(j,i)$
- Berechne Matrixelemente
$$M_x(i, j) = \log \left(\frac{f(i, j)}{f(i) * f(j)} \right)$$

Erläuterung

- Typische Formel für **Log-Odds-Ratio**
- Benutzung des Logarithmus zur Ersetzung von Multiplikation mit Addition
 - Normierung der Übergangshäufigkeit mit Wahrscheinlichkeit des zufälligen „Umkippens“

$$M_x(i, j) = \log \left(\frac{f(i, j)}{f(i) * f(j)} \right)$$

- $M(i, j) = 0$ (Bruch = 1)
 - **Keine Selektion** - Anzahl Übergänge entspricht statistischer Erwartung
- $M(i, j) < 0$ (Bruch < 1)
 - **Negative Selektion** – Übergang wird unterdrückt
- $M(i, j) > 0$ (Bruch > 1)
 - **Positive Selektion** – Übergang wird bevorzugt

Blosum Matrix

- Hauptkritikpunkte am PAM Ansatz
 - Nur Verwendung sehr ähnlicher Sequenzen
 - Realistische Zahlen für evolutionär weiter entfernte Sequenzen?
 - Einbeziehung kompletter Proteinsequenzen
 - Unterliegen die alle der selben Mutationsrate?
 - PAM-x vervielfältigen Fehler in PAM-1
- Anderer (neuerer) Ansatz: BLOSUM
 - **BLO**cks **SU**bstitution **M**atrix
 - **M**ultiples **A**lignments **e**volutionär **e**ntfernter, **a**ber **h**omologer **P**roteinsequenzen
 - Benutzung nur der konservierten Blöcke
 - Populärer als PAM Matrizen

BLOSUM – Vorarbeiten

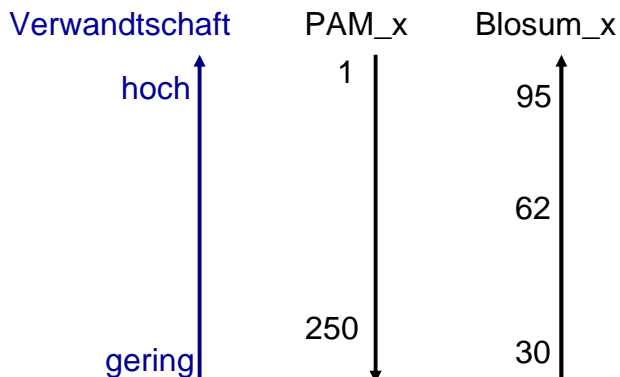
- PROSITE
 - Beschreibung funktionaler (=konservierter) Bereiche in [homologen Proteinsequenzen](#) durch reguläre Ausdrücke
 - Expertenwissen - manuelle Pflege der Datenbank am EBI
- BLOCKS
 - Alignierung von durch PROSITE Ausdrücke gematchten Sequenzen in [Multiple Alignments \(MSA\)](#)
 - Multiple Sequence Alignment
 - Heute: Verwendung weiterer Domänen aus PRINTS, PFAM, ...
 - Ein **BLOCK** ist zusammenhängendes Stück in einem MSA

```
FMYMFYVVPL_PQ_QVY
FYQQF_VQLYP_MFQV_
FMY_YUVQQP_UMUQ_
```

- Dann: Berechnung der Häufigkeiten nach PAM $M_x(i, j) = \log\left(\frac{f(i, j)}{f(i)*f(j)}\right)$

PAM vs. Blosum

- Zahlenwert bezieht sich auf Verwandtschaft von 2 Proteinen



Ähnlichkeit

- Welche Frage will man eigentlich beantworten?
 - Wie weit entfernt sind diese beiden Sequenzen
 - Wie ähnlich sind sich zwei Sequenzen
- Ähnlichkeit ist oft einfacher
 - Intuitives Maß – je ähnlicher, desto höher ist die Wahrscheinlichkeit für ähnliche Funktion
 - Programme berechnen i.d.R. Ähnlichkeit
 - Logisch ist Ähnlichkeit und Abstand äquivalent
- Differenzierte Betrachtung
 - Ähnlichkeit einzelner Zeichen / Basen / Aminosäuren
 - Ähnliche Zeichen – hohe positive Werte
 - Unähnliche Zeichen – negative Werte

Beispiel

$$\Sigma' = \{A, C, G, T, _\}$$

	A	C	G	T	_
A	4	-2	-2	-2	0
C		4	-2	-2	-2
G			4	-1	0
T				4	-2
_					0

A	C	_	G	T	C
A	G	G	T	_	C

= 3

A	C	G	T	C
A	G	G	T	C

= 14

Formal

- Definition

Gegeben Alphabet $S' = \Sigma' \cup _$, Strings A, B über S' mit $|A| = |B| = n$

- Eine **Scoringfunktion** ist eine Funktion $s: S' \times S' \rightarrow \mathbb{Z}$
 - Substitutionsmatrix
- Die **Ähnlichkeit** von A, B bzgl. der Scoringfunktion s ist

$$\text{sim}(A, B) = \sum_{i=1}^n s(A[i], B[i])$$

- Bemerkung

- Optimales Alignment ~ Alignment mit höchster Ähnlichkeit

Rekursionsgleichung

- Nur kleine Veränderung

$$d(i, 0) = \sum_{i=1}^n s(A[i], _) \quad d(0, j) = \sum_{i=1}^m s(_, B[i])$$

$$d(i, j) = \max \left\{ \begin{array}{l} d(i, j-1) + s(_, B[j]) \\ d(i-1, j) + s(A[i], _) \\ d(i-1, j-1) + s(A[i], B[j]) \end{array} \right\}$$

Begriffe

- Logisch äquivalent:
 - Alignment
 - Untereinanderordnung von Strings
 - Einfügen von Spaces
 - Alignmentsscore zählt Mismatches und Spaces
 - Alignmentabstand: Minimiere Alignmentsscore
 - Editskripte
 - Operationen auf Zeichen
 - Editabstand zählt Operationen R, I, D
 - Dynamische Programmierung: **Minimiert** Editabstand
 - Ähnlichkeit
 - Differenzierte Betrachtung der „Ähnlichkeit“ von Zeichen
 - Ähnlichkeitsmaß summiert Zeichenähnlichkeiten
 - Dynamische Programmierung: **Maximiert** Ähnlichkeit
-

Aufgabe 4

- **Ähnlichkeitsmaß** von 2 Proteinsequenzen bestimmen
 - Dynamische Programmierung mit Substitutionsmatrix
 - **Alignment** ausgeben
 - besten Alignments der beiden String ausgeben
-

Eingabe

■ Einfach

- 2 Strings oder Dateien innerhalb der Anwendung definieren
- Datei der Substitutionsmatrix in der Anwendung definieren

■ Erweitert

- Dateien über args[] eingeben

Daten aus Datei auslesen

Insert

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

Leerzeichen

pos. / neg. Integerwerte

Blosum62

Daten aus Datei auslesen

- Auslesen
 - OutputStream
 - Tokenizer
- Speichern
 - Aminosäuren in char-Array
 - Werte in einem Integer-Array

Ausgabe



- **Tabelle** der berechneten Ähnlichkeitswerte ausgeben
 - mit Sequenzen
- **Mögliche Alignments**
 - einfach: nur die Alignments
 - komplex: mit Consensus Sequenz
 - + für ähnliche (z.B. E – K)
 - Buchstaben für gleiche

Sequenz1	SIGGYD
cons Seq	S+GG
Sequenz2	SVGGNR

Sequenzen

- Testsequenzen
 - test_proteins.txt ? eine Tabelle auf Blatt berechnen
- Biologische Daten
 - PH4H_ - Phenylalanine hydroxylase
 - CAEEL Caenorhabditis elegans (Fadenwurm)
 - CAUCR Caulobacter crescentus (Bakterium)
 - CHRVO Chromobacterium violaceum (Bakterium)
 - DROME Drosophila melanogaster (Fruchtfliege)
 - HUMAN Homo sapiens (Mensch)
 - MOUSE Mus musculus (Maus)
 - RAT Rattus norvegicus (Ratte)
 - VIBCH Vibrio cholerae (Bakterium)