

Bioinformatik

Substitutionsmatrizen
BLAST

Ulf Leser

Wissensmanagement in der
Bioinformatik

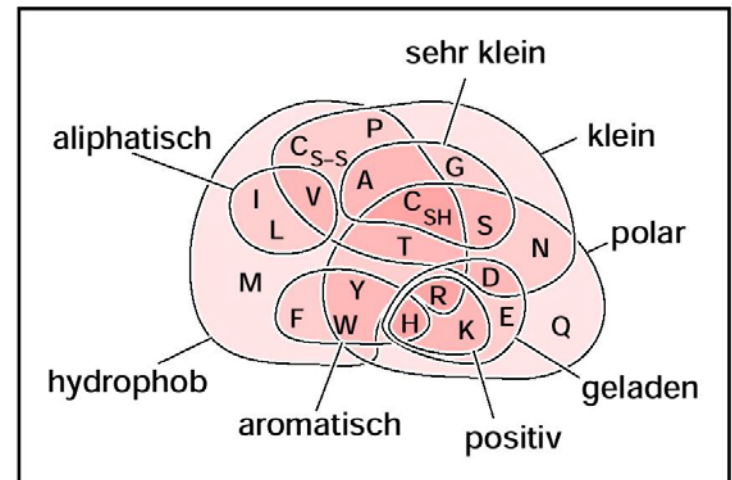


Inhalt dieser Vorlesung

- Substitutionsmatrizen: PAM und BLOSSUM
- Suche in Datenbanken: Basic Local Alignment Search Tool (BLAST)

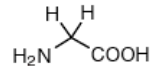
Hintergrund

- Schon öfters angesprochen ...
 - Ähnlichkeitsmatrizen, Substitutionsmatrizen, Scorefunktionen, ...
- Ersetzung einer Base/Aminosäure durch eine andere hat **unterschiedliche Bedeutung**
 - Basen: Auswirkungen auf kodiertes 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

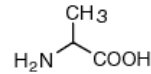


Unterschiedliche Aminosäuren

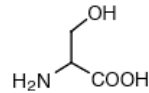
Small



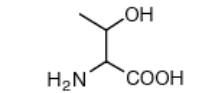
Glycine (Gly, G)
MW: 57.05



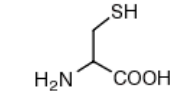
Alanine (Ala, A)
MW: 71.09



Serine (Ser, S)
MW: 87.08, pK_a ~ 16

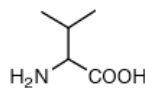


Threonine (Thr, T)
MW: 101.11, pK_a ~ 16

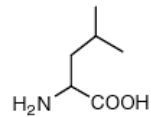


Cysteine (Cys, C)
MW: 103.15, pK_a = 8.35

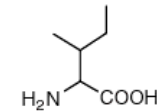
Hydrophobic



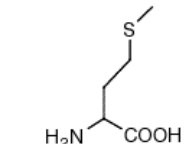
Valine (Val, V)
MW: 99.14



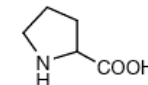
Leucine (Leu, L)
MW: 113.16



Isoleucine (Ile, I)
MW: 113.16

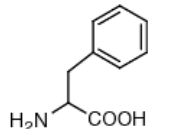


Methionine (Met, M)
MW: 131.19

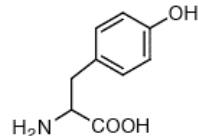


Proline (Pro, P)
MW: 97.12

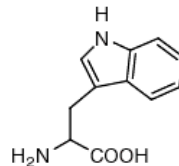
Aromatic



Phenylalanine (Phe, F)
MW: 147.18

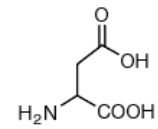


Tyrosine (Tyr, Y)
MW: 163.18

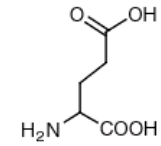


Tryptophan (Trp, W)
MW: 186.21

Acidic

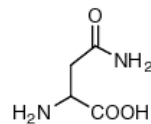


Aspartic Acid (Asp, D)
MW: 115.09, pK_a = 3.9

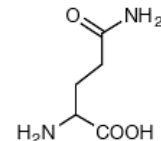


Glutamic Acid (Glu, E)
MW: 129.12, pK_a = 4.07

Amide

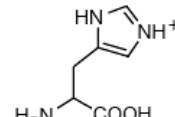


Asparagine (Asn, N)
MW: 114.11

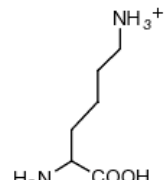


Glutamine (Gln, Q)
MW: 128.14

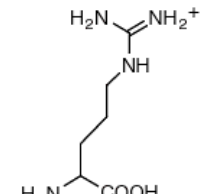
Basic



Histidine (His, H)
MW: 137.14, pK_a = 6.04



Lysine (Lys, K)
MW: 128.17, pK_a = 10.79



Arginine (Arg, R)
MW: 156.19, pK_a = 12.48

Ist das alles notwendig?

Code	Häufigkeit	Mutierbarkeit
L	0.091	54
A	0.077	100
G	0.074	50
S	0.069	117
V	0.066	98
E	0.062	77
K	0.059	72
T	0.059	107
I	0.053	103
D	0.052	86
P	0.051	58
R	0.051	83
N	0.043	104
Q	0.041	84
F	0.040	51
Y	0.032	50
M	0.024	93
H	0.023	91
C	0.020	44
W	0.014	25

- Häufigkeiten der einzelnen Aminosäuren im Verhältnis zur Gesamtzahl
- Häufigkeiten der Ersetzung einer Aminosäure im Verhältnis zu allen Ersetzungen
- Alanin (A) willkürlich als 100% gesetzt
- **Keinesfalls Gleichverteilung**
- Es gibt klar bevorzugte Mutationen
 - Besser: durch Selektion klar benachteiligte Mutationen
 - Tryptophan (W) sehr selten (25)
 - Serin (S) sehr häufig (117)

Substitutionsmatrizen

- Bewertung **individueller Substitutionen**
- Alignmentalgorithmus ist davon unberührt, nur Parameter ändern sich
- Das Ergebnis kann sich aber **vollkommen ändern**
- Beispiele: Blosom62, Identitätsmatrix

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

	A	C	G	T
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5

Woher nehmen?

- Wie kann man sinnvolle Werte für die Matrix bestimmen?
 - Wir wollen **Ähnlichkeit der biologischen Bedeutung** messen
- Möglichkeit 1: Chemische Eigenschaften
 - Ladung, Größe, Polarität, ...
 - Viele Faktoren mit unklaren Gewichten
 - Wie soll man das durch **ein Bewertungsschema** ausdrücken?
 - Keine Verwendung in der Praxis
- Möglichkeit 2: Beobachtung
 - **Beobachtung der Evolution** statt analytischer Vorhersage
 - Lernen aus Beispielen, also „tatsächlich“ vorgekommener Mutationen
 - Benötigt große Menge homologe Sequenzen

PAM: Point-Accepted Mutations

- Dayhoff et al., „A model of evolutionary change in proteins“, 1978
- PAM: Zwei Bedeutungen
 - 1 PAM – **Einheit** für den Abstand von Proteinsequenzen
 - PAM-X Matrix – Berechnete **Substitutionsmatrix** für zwei Sequenzen die X PAM entfernt sind

1. PAM als Sequenzabstand

- Definition

Seien S_1 und S_2 zwei Proteinsequenzen. S_1 und S_2 heißen x PAM entfernt, wenn gilt

- *S_1 wurde in S_2 überführt mit x Punktmutationen pro 100 Aminosäuren*

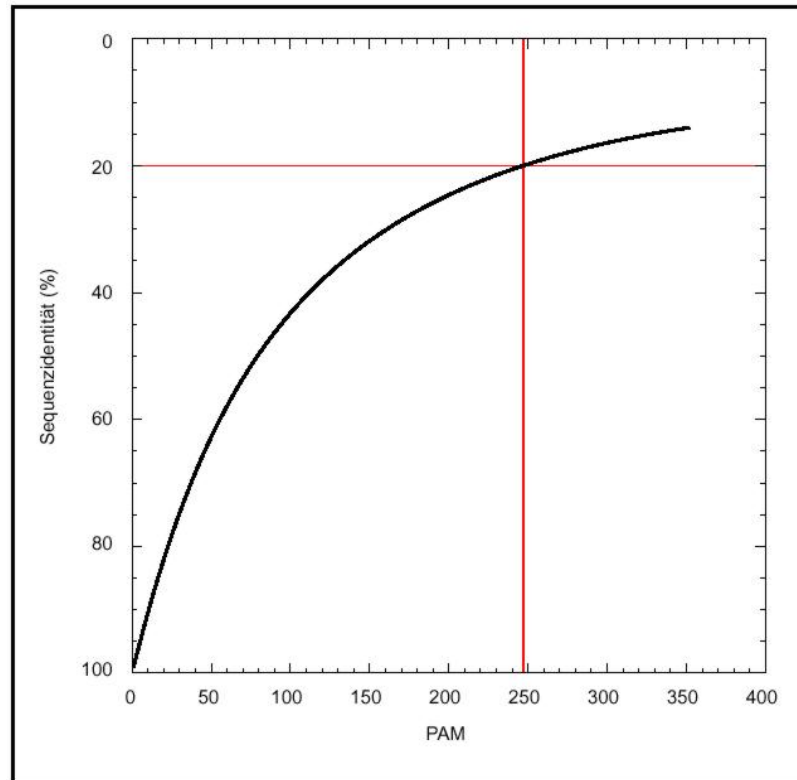
- Eigenschaften

- PAM beachtet keine Inserts und Deletions
 - Großes Manko des PAM Ansatzes
- Der echten PAM Abstand zweier Sequenzen ist nicht bestimmbar
 - Dazu hätte man ein paar Millionen Jahre zusehen müssen
- Stattdessen: **Schätzen des PAM Abstandes**
 - Grundlage: durchschnittliche Veränderung nach X Mutationen
 - Ermitteltbar z.B. durch Simulation
- Vorsicht: 50 PAM heißt nicht 50 veränderte pro 100 Aminosäuren
 - Doppelmutationen, Rückmutationen, etc.

- Eigentlich besser

- *S_1 , S_2 sind x PAM entfernt, wenn: S_1 wurde **am wahrscheinlichsten** in S_2 überführt mit x Punktmutationen pro 100 Aminosäuren*

PAM Abstand und Sequenzidentität



- Jenseits von PAM 250 ist nur noch Rauschen

2. PAM Matrizen - Grundidee

- Vorgehen

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 PAM- x Matrix M_x wie folgt

- Messe absolute Häufigkeit $f(A_i)$ für alle Aminosäuren A_i über alle Sequenzen, normiert auf Gesamtlänge aller Sequenzen
- Aligniere alle Paare entsprechend der **evolutionären Wahrheit**
 - $S_{k,l}'$ sei die Sequenz $S_{k,l}$ mit den durch das Alignment eingefügten Leerzeichen
- Messe **Übergangshäufigkeiten $f(i,j)$** zwischen allen Paaren von Aminosäuren (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
- Bruch
 - Normierung der Übergangshäufigkeit mit Wahrscheinlichkeit des zufälligen „Umkippens“

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

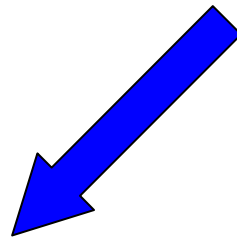
- $M(i, j) = 0$ (Bruch = 1)
 - **Keine Selektion** - Anzahl Übergänge entspricht statistischer Erwartung
- $M(i, j) < 0$ (Bruch < 1)
 - **Negative Selektion** – Übergang wird unterdrückt
- $M(i, j) > 0$ (Bruch > 1)
 - **Positive Selektion** – Übergang wird bevorzugt

Beispiel

$S_{1,1}$: ACGGTGAC
 $S_{2,1}$: AGG_TGCC
 $S_{1,3}$: GTT_AGCTA
 $S_{2,4}$: TTTCAG_TA
 $S_{1,2}$: GGTCAA
 $S_{2,2}$: AGTC_A

Absolute Häufigkeiten

A: 11/42	C: 8/42	G: 12/42	T: 11/42
----------	---------	----------	----------



Übergangshäufigkeiten



	A	C	G	T
A	4/19	1/19	1/19	0/19
C		2/19	1/19	0/19
G			4/19	1/19
T				5/19

Substitutionsmatrix

	A	C	G	T
A	0,48	0,02	-0,15	-
C		0,46	-0,01	-
G			0,41	-0,15
T				0,58

Probleme mit PAM Matrizen

- Es gibt keinen Algorithmus, der diese Matrizen berechnet
 - „**Evolutionäre Wahrheit**“ ist nicht bekannt
 - Besonders bei großem Abstand ist Bestimmung eines „wahren“ Alignments sehr schwierig und subjektiv

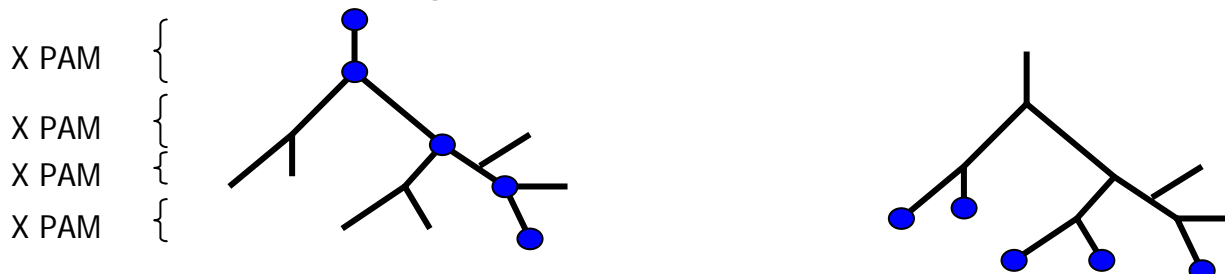
einfach

```
FMMIYVVYL   FMM_IYVVYL
FMMUIYVVYL   FMMUIYV_YL
```

schwierig

```
_FMMFYVVYL
UFPHVYL_YQ
FMMFYVVYL
UFPHVYLYL
FMMFYVVYL__
__UFPHVYLYL
```

- Wir haben **keine Sequenzen, die x PAM entfernt** sind
 - Die „wahre“ Entfernung kann man immer nur schätzen



Reale PAM Matrizen

- Vorgehen von Dayhoff et al.
 - Paare eng verwandter Sequenzen auswählen
 - >85% Identität, 34 Proteinfamilien
 - Manuell alignieren
 - PAM-1 Matrix M_1 aus Häufigkeiten berechnen
 - PAM-x Matrizen wie folgt berechnen: $M_n = (M_1)^n$
- Dem liegen viele Annahmen zugrunde
 - **Evolutionary Clock Theory**: Evolution verläuft gleichmäßig
 - in der Zeit und in den Sequenzpositionen
 - Proportionalität von Veränderungen
 - Hochrechnung langer Distanzen aus kurzen
 - Keine Insertions oder Deletions
 - Unabhängigkeit der Mutationswahrscheinlichkeit von der Position in der Sequenz (und der Nachbarschaft)

Verwendung

- Welche PAM Matrix soll man nun zur Alignierung zweier Sequenzen verwenden?
 - Die, die dem PAM-Abstand der Sequenzen entspricht
 - Den kennt man aber nicht – schätzen
 - Schätzung benötigt Alignments
 - Zur Berechnung der Sequenzidentität
 - Alignments basieren auf Substitutionsmatrizen
 - **Henne – Ei Problem**
- Also
 - Verschiedene Matrizen testen
 - Alignments prüfen
 - Externes Wissen (Chemie, Strukturen, etc.) hinzuziehen

BLOSUM Matrizen

- Hauptkritikpunkte am PAM Ansatz
 - Nur Verwendung sehr ähnlicher Sequenzen
 - Realistische Zahlen für evolutionär weiter entfernte Sequenzen?
 - Manuelles Alignment
 - Stimmt das?
 - PAM-x für hohe x vervielfältigen eventuelle Fehler in PAM-1
 - Schwierigkeit, das korrekt „x“ für konkrete Alignmentaufgaben auszuwählen
- Anderer (neuerer) Ansatz: BLOSUM Matrizen
 - **B**LOcks **S**Ubstitution **M**atrix
 - Basiert auf **multiplen Alignments evolutionär entfernter, aber homologer Proteinsequenzen**
 - Populärer als PAM Matrizen

BLOSUM Vorarbeiten

- PROSITE

- Beschreibung identifizierender (funktionstragender?) Bereiche in **homologen Proteinsequenzen** durch reguläre Ausdrücke
- Expertenwissen - manuelle Pflege der Datenbank am EBI

- BLOCKS

- Alignierung der durch PROSITE Ausdrücke gematchten Sequenzen in **Multiple Alignments** (MSA)
 - Multiple Sequence Alignment – Dazu später mehr
 - BLOSUM heute: Verwendung weiterer Domänen aus PRINTS, PFAM, ...
- Ein BLOCK ist zusammenhängendes Stück in einem MSA

```
FMYMFYVVPL_PQ_QVY
FYQQF_VQLYP_MFQV_
FMY_YUVOQP_UMUQ_
```

BLOSUM Matrizen

- Berechnung der BLOSUM Matrizen verläuft identisch zur Berechnung der PAM-1 Matrix
 - Alle BLOCKS werden betrachtet
 - Absolute Häufigkeiten aller Aminosäuren
 - Häufigkeiten aller Übergänge in allen Paaren

$$M_1(i, j) = \log\left(\frac{f(i, j)}{f(i) * f(j)}\right)$$

- BLOSUM-x Matrizen
 - Bias in BLOCKS durch zu viele sehr ähnliche Sequenzen
 - Zur Berechnung der BLOSUM-x Matrix werden in jedem Block alle Sequenzen mit >x% Identität zu einer Sequenz zusammengefasst
 - **Gänzlich andere Bedeutung** des „x“ als in PAM-x
 - Aber ähnliche Verwendung: x ~ evolutionärem Abstand

Unterschiede PAM - BLOSUM

- BLOSUM verwendet nur hochkonservierte Bereiche, PAM komplette Alignments
- PAM rechnet große evolutionäre Abstände nur hoch, BLOSUM verwendet gezielt entfernte Sequenzen
- BLOSUM basiert auf deutlich mehr Sequenzen
- Heutige BLOSUM-Matrizen sind heuristisch verbessert
 - „Feedback-Schleife“: Mit initialer BLOSUM 62 Matrix erneute Alignierung
 - Bestimmung der BLOCKS verwendet BLOSUM Matrix
- Hochgradig heuristisches Feld
- BLOSUM-62 oft Default in Alignmentprogrammen