

# Algorithmische Bioinformatik

## Substitutionsmatrizen

Ulf Leser

Wissensmanagement in der  
Bioinformatik



# Ziele

---

- Kenntnis einer typischen realen bioinformatischen Vorgehensweise: Viel schätzen, viel ignorieren, viel abkürzen (und es funktioniert doch halbwegs)

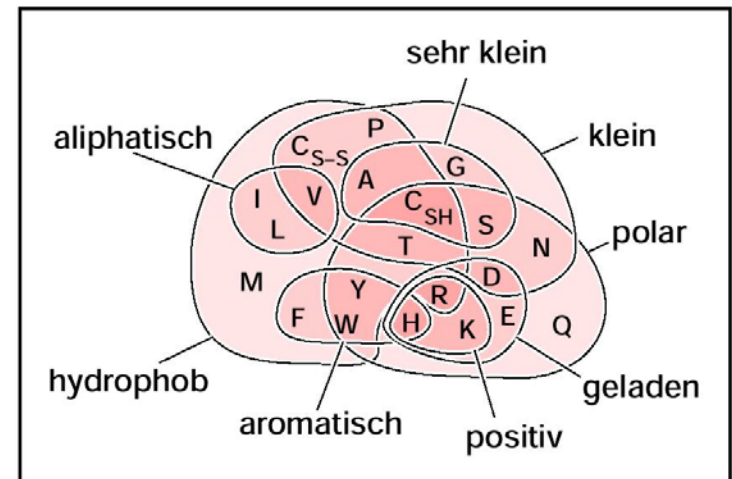
# Inhalt dieser Vorlesung

---

- Mutationswahrscheinlichkeiten
- Jukes-Cantor Modell (für DNA)
- Echte Substitutionsmatrizen

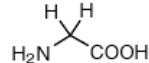
# Hintergrund

- Warum Ähnlichkeitsmatrizen, Substitutionsmatrizen, ...?
- Ersetzung einer Aminosäure durch eine andere hat **unterschiedliche Bedeutung** je nach Paar
  - Ersetzung mit sehr ähnlichen Aminosäuren ändert Proteinstruktur in der Regel kaum
  - Ersetzung mit wenig ähnlichen Aminosäuren kann Struktur vollkommen ändern

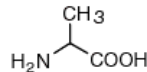


# Unterschiedliche Aminosäuren

## Small

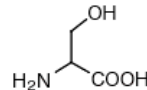


Glycine (Gly, G)  
MW: 57.05

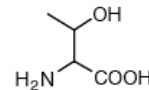


Alanine (Ala, A)  
MW: 71.09

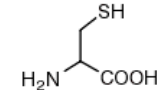
## Nucleophilic



Serine (Ser, S)  
MW: 87.08,  $pK_a \sim 16$

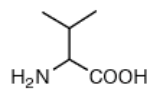


Threonine (Thr, T)  
MW: 101.11,  $pK_a \sim 16$

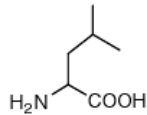


Cysteine (Cys, C)  
MW: 103.15,  $pK_a = 8.35$

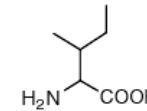
## Hydrophobic



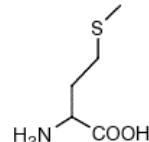
Valine (Val, V)  
MW: 99.14



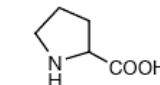
Leucine (Leu, L)  
MW: 113.16



Isoleucine (Ile, I)  
MW: 113.16

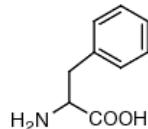


Methionine (Met, M)  
MW: 131.19

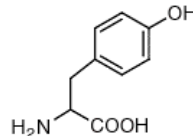


Proline (Pro, P)  
MW: 97.12

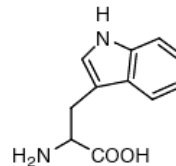
## Aromatic



Phenylalanine (Phe, F)  
MW: 147.18

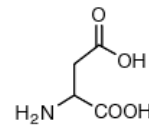


Tyrosine (Tyr, Y)  
MW: 163.18

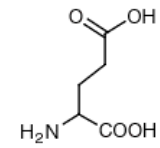


Tryptophan (Trp, W)  
MW: 186.21

## Acidic

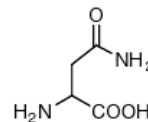


Aspartic Acid (Asp, D)  
MW: 115.09,  $pK_a = 3.9$

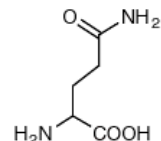


Glutamic Acid (Glu, E)  
MW: 129.12,  $pK_a = 4.07$

## Amide

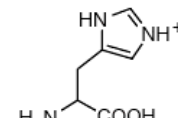


Asparagine (Asn, N)  
MW: 114.11

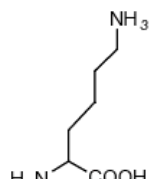


Glutamine (Gln, Q)  
MW: 128.14

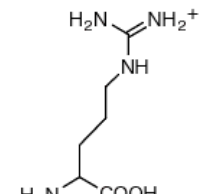
## Basic



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,  $pK_a = 12.48$

# Substitutionsmatrizen

- Individuelle Bewertung von Mismatches und Matches

Blosum62

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

DNA-Matrix von BLAST

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

# Ist das 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 Ersetzung** einer Aminosäure durch irgendeine andere AA im Verhältnis zu allen Ersetzungen
- Alanin (A) willkürlich als 100% gesetzt
- **Keine Gleichverteilung**
- Mutationen sind mehr oder weniger erfolgreich, je nachdem, welche AA ersetzt wird
  - Besser: Mutationen werden durch Selektion mehr oder weniger geduldet
  - Tryptophan (W) sehr selten (25)
  - Serin (S) sehr häufig (117)

# Woher nehmen?

---

- Für Aminosäuren benötigt man ~200 (verschiedene) Werte
- Wie kann man sinnvolle Werte für die Matrix bestimmen?
- Möglichkeit 1: Physikalische / 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: Empirisch
  - Beobachtung der Evolution statt analytischer Vorhersage
  - Lernen aus Beispielen: Tatsächlich vorgekommener Mutationen
  - Benötigt Beispieldaten, also Paare von homologen Sequenzen



# Inhalt dieser Vorlesung

---

- Mutationswahrscheinlichkeiten
- Jukes-Cantor Modell (für DNA)
- Echte Substitutionsmatrizen

# Evolutionäre Abstände

---

- Wichtige Unterscheidung
  - Mutation: Evolutionäres Ereignis, **nicht beobachtbar**
  - Substitution: **Beobachteter Mismatch**; Mutation, die überdauert hat
- Gesucht: Maß für evolutionären Abstand zweier Sequenzen
- Populär: Menge an **stattgefundenen Punktmutationen**
  - Ignoriert InDels (selten in kodierenden Bereichen)
  - Nimmt Gleichverteilung von Mutationen in Zeit und Raum an
- Messen kann man aber nur Zahl der **Substitutionen**
- Wie **viele Mutationen sind im Schnitt notwendig**, um in einer  $n$  Zeichen langen Sequenz  **$x$  Substitutionen** zu erzeugen?

# Jukes-Cantor-Modell

---

- Lösung hängt vom **Evolutionsmodell** ab
  - Wahrscheinlichkeiten aller Basenmutationen
  - Abhängigkeiten von Mutationen zwischen verschiedenen Positionen
  - Abhängigkeit der Wahrscheinlichkeiten von Sequenzlänge
  - ...
- Einfachstes Modell: **Jukes-Cantor**
  - Jukes, Cantor (1969). *Evolution of Protein Molecules*. Academic Press
  - Alle Positionen werden mit der gleichen Wsk verändert
  - Alle Basenmutationen sind gleichwahrscheinlich
  - Wahrscheinlichkeiten ändern sich nicht über die Zeit
- Sei  $\alpha$  die Wsk des Auftretens einer **beliebigen Mutation** an einer festen Position

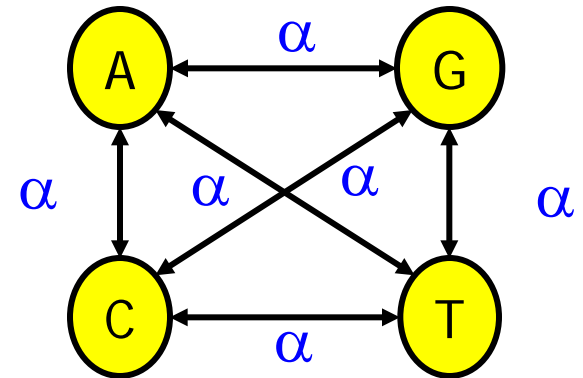
# Ableitung [EG01]

- Sei  $p_{CC}(t)$  die Wsk, dass an **einer geg. Position nach  $t$  Mutationen** ein C steht, wenn dort auch im Original ein C steht. Dann gilt:
  - $p_{CC}(1) = 1 - (CA \vee CG \vee CT) = 1 - 3 \cdot \alpha$
  - $p_{CC}(2) = p_{CC}(1) \cdot (1 - 3\alpha) + (1 - p_{CC}(1)) \alpha$
  - $p_{CC}(t+1) = p_{CC}(t) \cdot (1 - 3\alpha) + (1 - p_{CC}(t)) \alpha$
- Auflösen ergibt (siehe Literatur,  $C \neq X$ )

$$p_{CC}(t) = \frac{1}{4} + \left(\frac{3}{4}\right) \cdot e^{-4\alpha t}$$

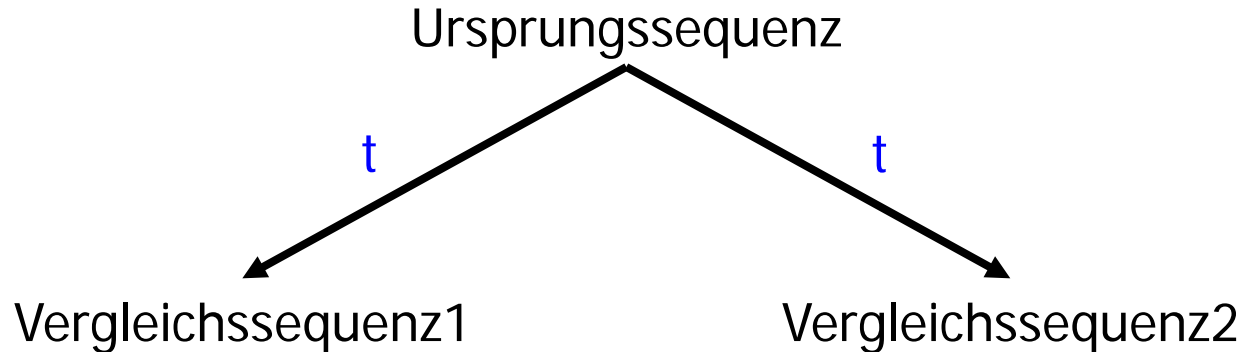
$$p_{CX}(t) = \frac{1}{4} - \left(\frac{1}{4}\right) \cdot e^{-4\alpha t}$$

- Bemerkung
  - Konvergiert beides gegen  $1/4$  , also gegen eine zufällige Sequenz



# Evolutionäre Verhältnisse

---



- Wsk, dass eine gegebene Base X aus der Ursprungssequenz **in beiden Ästen nach t Schritten nicht mutiert** ist:

$$p_{XX}(t) * p_{XX}(t) = p_{XX}(t)^2$$

- Wsk, dass eine gegebene Base X aus der Ursprungssequenz **in beiden Ästen nicht substituiert** ist:

$$I(t) = p_{XX}(t)^2 + \sum_{B \neq X} P_{BX}(t)^2 = \frac{1}{4} + \frac{3}{4} e^{-8t\alpha}$$

# Von Substitutionen zu Mutationen

---

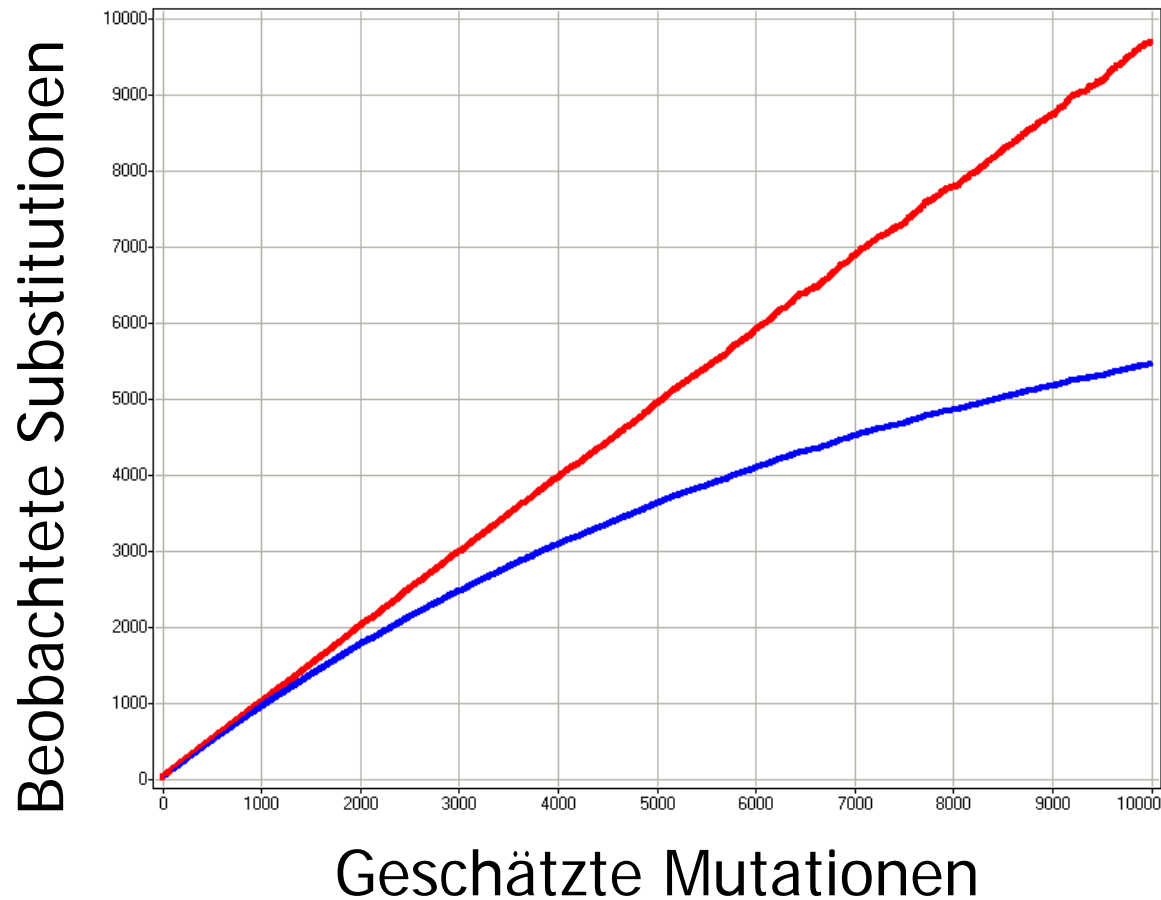
- Wsk, dass eine beliebige Base **substituiert ist**:

$$p(t) = 1 - I(t) = 3/4 - 3/4 * e^{-8\alpha t} = 3/4 * (1 - e^{-8\alpha t})$$

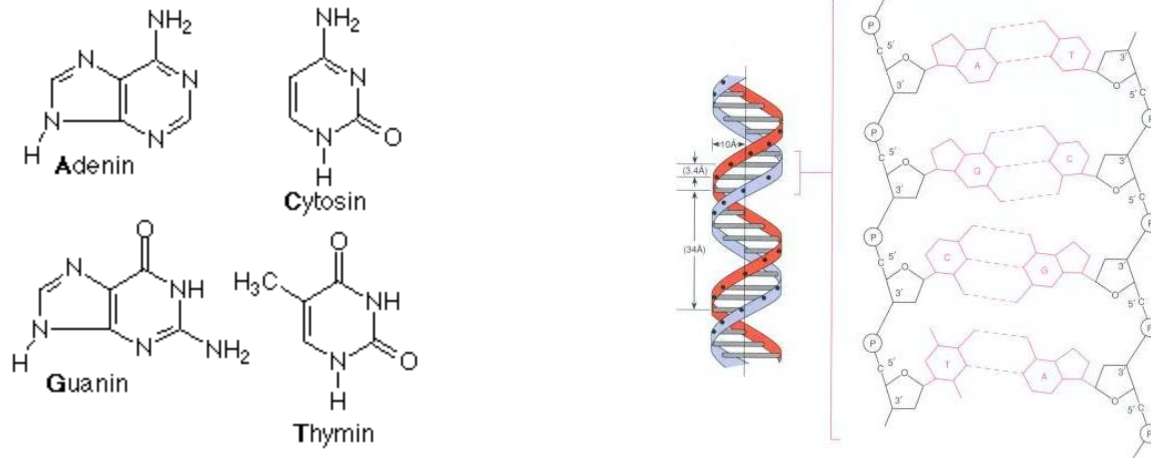
- Umformung ergibt:  $8\alpha t = -\ln(1 - 4/3p(t))$
- Wir wollen wissen, für wie viele Mutationen der Erwartungswert der entstehenden Substitutionen gleich der Zahl beobachteter Substitutionen ist
  - Dann können wir aus der Anzahl beobachteter Substitutionen die wahrscheinlichste Anzahl tatsächlicher Mutationen schätzen
  - Wie viele **Mutationen** gab es? Im Schnitt  $k = 3\alpha * 2t$
  - Umformen und einsetzen für  $\alpha$  ergibt:  $8t * k/6/t = -\ln(1 - 4/3p(t))$
  - Zusammen:  $k = -3/4 * \ln(1 - 4/3p(t))$ 
    - $p(t)$  ist der **Erwartungswert, dessen empirischen Wert** wir kennen: Zahl der Substitutionen = %-Diversität
  - [Ausführliche Ableitung: <http://www.montefiore.ulg.ac.be/~kvansteen/GBIO0009-1/ac20132014/T4/jc.pdf>]

# Jukes-Cantor Korrektur ( $|P|=10.000$ )

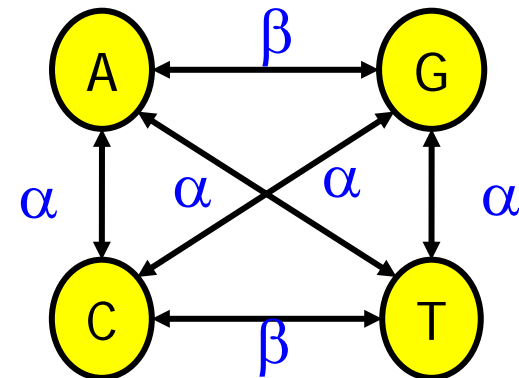
---



# Kimura's Modell



- Pyrimidine: Cytosin, Thymin, Uracil: Ein Ring
- Purine: Adenin, Guanin: Zwei Ringe
- **Transversion**: Purin  $\Leftrightarrow$  Pyrimidin oder umgekehrt
- **Transition**: Purin  $\Leftrightarrow$  Purin oder Pyrimidin  $\Leftrightarrow$  Pyrimidin
- Damit:  $k = 1/2 \ln[1/(1 - 2P - Q)] + 1/4 \ln[1/(1 - 2Q)]$ 
  - P: %-Anteil Transitionen, Q: %-Anteil Transversionen





# Inhalt dieser Vorlesung

---

- Mutationswahrscheinlichkeiten
- Jukes-Cantor Modell (für DNA)
- Echte Substitutionsmatrizen
  - PAM: Point-Accepted Mutations
  - BLOSUM: Blocks Substitution Matrices

# Margaret O. Dayhoff

- “Deduce **evolutionary relationships** of the biological kingdoms, phyla, and other taxa from sequence evidence”
- Collection of all known protein sequences
- First edition: 65 proteins
- Several **printed releases**
- Resulted in the Protein Information Resource (PIR)

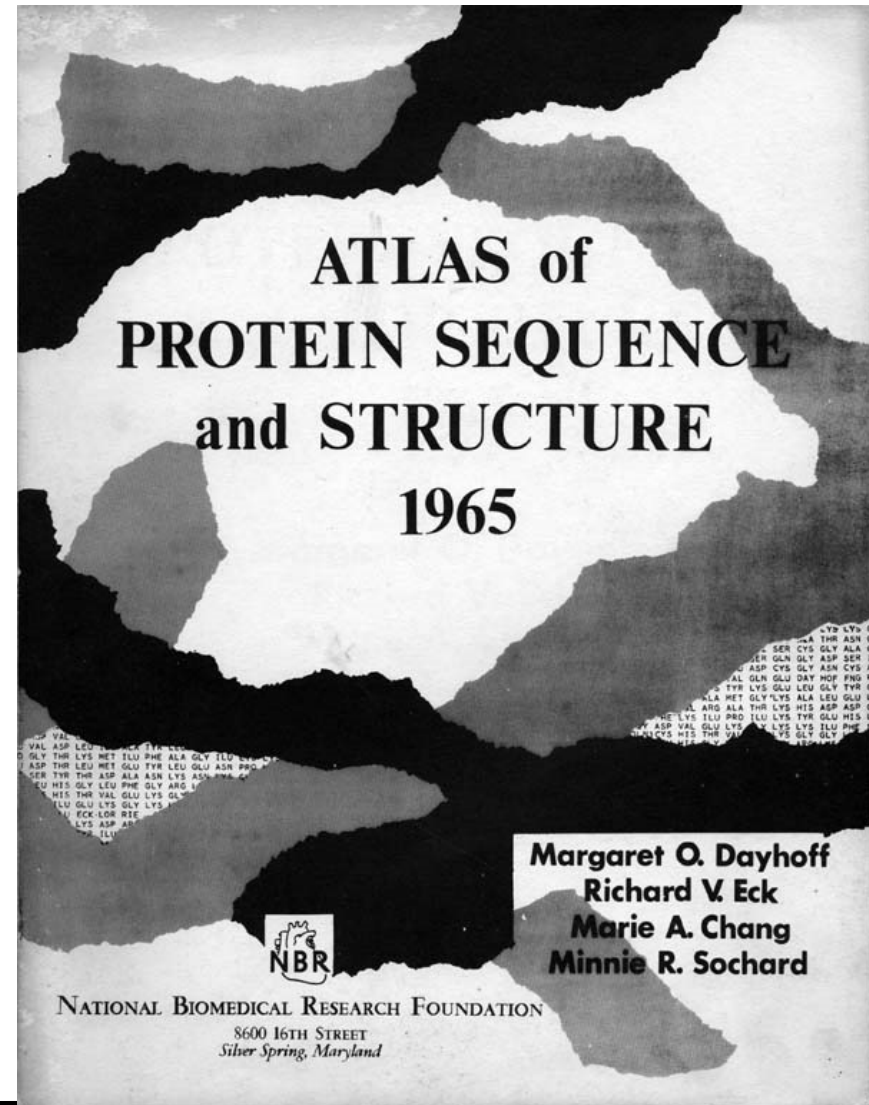


Bild: Dank an Antje Krause

# PAM: Point-Accepted Mutations

---

- Dayhoff MO, Schwartz RM. *A model of evolutionary change in proteins*. Atlas of protein sequence and structure 1978.
- Zwei Bedeutungen
  - 1 PAM – **Einheit** für den evolutionären Abstand von Proteinsequenzen
  - PAM-X Matrix – Berechnete **Substitutionsmatrix** für zwei Sequenzen, die X PAM entfernt sind

# PAM als Sequenzabstand

---

- Definition

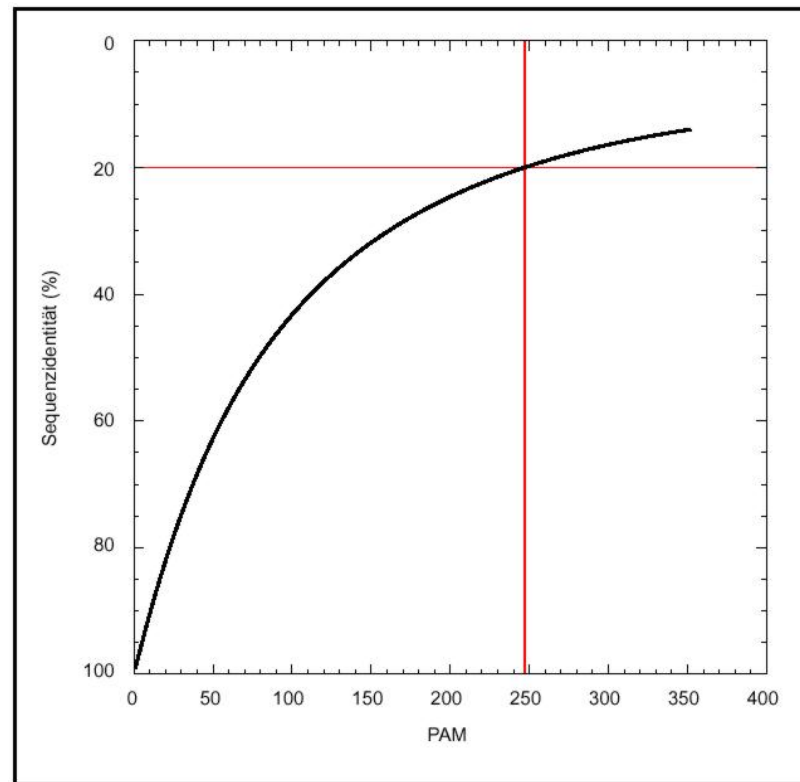
*Seien  $S_1$  und  $S_2$  Proteinsequenzen mit  $|S_1|=|S_2|$ .  $S_1$  und  $S_2$  heißen  **$x$  PAM entfernt**, wenn  $S_1$  in  $S_2$  wahrscheinlich mit  **$x$  Punktmutationen pro 100 Aminosäuren** überführt wurde*

- Eigenschaften

- PAM beachtet keine Inserts und Deletions
- $x$  schätzt man aus den Substitutionen zwischen  $S_1$  und  $S_2$ 
  - Berechnung ähnlich Jukes-Cantor, aber auf Aminosäuren
- 50 PAM Abstand heißt also nicht etwa 50 Substitutionen pro 100 AA

# PAM Abstand und Sequenzidentität

---



- Jenseits von PAM 250: Rauschen

# PAM Matrizen

---

- Definition: Seien  $(S_{1,1}, S_{2,1}), \dots, (S_{1,n}, S_{2,n})$  Paare von AA-Sequenzen, die jeweils  $x$  PAM entfernt sind. Dann ist die **PAM- $x$  Matrix  $M_x$**  wie folgt definiert:
  - Sei  $f(i)$  die relative Häufigkeit der Aminosäure  $A_i$
  - Seien alle Paare optimal aligniert
    - Sei  $S_{k,l}'$  das  $S_{k,l}$  mit den durch das Alignment eingefügten Leerzeichen
  - Sei  **$f(i,j)$  die relative Übergangshäufigkeit** von  $A_i$  zu  $A_j$  in allen alignierten Paaren
    - Relativer Anteil von Positionen  $k$  mit  $S_{1,z}'[k]=A_i$  und  $S_{2,z}'[k]=A_j$  oder umgekehrt
      - Übergang ist „richtungslos“:  $f(i,j) = f(j,i)$
      - Paare  $(A_x, \_)$  werden ignoriert
  - Damit:
$$M_x(i, j) = \log \left( \frac{f(i, j)}{f(i) * f(j)} \right)$$

# Erläuterung

---

- Log-Odds Ratio

- Numerischer Trick: Logarithmus zur Ersetzung von Mult durch Add

- Bedeutung des Bruchs

- Verhältnis der beobachteten Substitutionen zu den erwarteten Substitutionen bei statistischer Unabhängigkeit

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

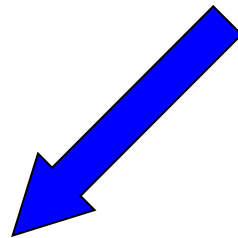
- $M(i, j) = 0$  (Bruch = 1)
  - Keine Selektion - Anzahl Übergänge entspricht statistischer Erwartung
- $M(i, j) < 0$  (Bruch < 1)
  - Negative Selektion – Übergang wird unterdrückt
- $M(i, j) > 0$  (Bruch > 1)
  - Positive Selektion – Übergang wird bevorzugt

# Beispiel

$S_{1,1}$ : ACGGTGAC  
 $S_{2,1}$ : AGG\_TGCC  
 $S_{1,2}$ : GTT\_AGCTA  
 $S_{2,2}$ : TTTCAG\_TA  
 $S_{1,3}$ : GGTC\_AA  
 $S_{2,3}$ : AGTC\_A

Relative 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



# Fehlerquellen

---

- Es kann **viele optimale Alignments** geben
  - Wahl eines Alignments besonders bei großem Abstand (also auch großen x) schwierig und subjektiv

Einfach: **FMM\_IYVVYL**  
**FMMUIYV\_YL**

Schwierig: **FMMF\_YV\_VYL**  
**\_\_UFPHVYLYL**

**\_FMMFYVVYL**  
**UFPHVYL\_YL**  
**FMMFYVVYL\_\_**  
**\_\_UFPHVYLYL**

- PAM Abstand ist nur eine Schätzung
  - Je ungenauer, je größer x ist
- Für größere Abstände benötigt ein gutes Alignment individuelle Mutationswahrscheinlichkeiten
  - Die suchen wir aber gerade

# Reale PAM Matrizen

---

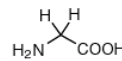
- Annahme der **Molecular Clock Theory**
  - Evolution verläuft gleichmäßig in der Zeit und in den Sequenzpositionen
- Damit: Hochrechnung der Matrixeinträge für lange  $x$  aus der Matrix für kurze  $x$
- Vorgehen von Dayhoff et al.
  - Paare eng verwandter Sequenzen auswählen
    - Sehr hohe %-Identität, 34 Proteinfamilien
  - Manuell alignieren
  - PAM-1 Matrix  $M_1$  aus Häufigkeiten berechnen
  - PAM- $x$  Matrizen berechnen durch:  $M_x = (M_1)^x$

# PAM 250

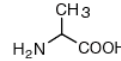
## 250 Multiplikationen der PAM-1 Matrix mit sich selber

Cys	12								
Gly	-3	5							
Pro	-3	-1	6						
Ser	0	1	1	1					
Ala	-2	1	1	1	2				
Thr	-2	0	0	1	1	3			
Asp	-5	1	-1	0	0	0	4		
Glu	-5	0	-1	0	0	0	3	4	
Asn	-4	0	-1	1	0	0	2	1	
Gln	-5	-1	0	-1	0	-1	2	2	
His	-3	-2	0	-1	-1	-1	1	1	
Lys	-5	-2	-1	0	-1	0	0	0	
Arg	-4	-3	0	0	-2	-1	-1	-1	
Val	-2	-1	-1	-1	0	0	-2	-2	
Met	-5	-3	-2	-2	-1	-1	-3	-2	
Ile	-2	-3	-2	-1	-1	0	-2	-2	
Leu	-6	-4	-3	-3	-2	-2	-4	-3	
Phe	-4	-5	-5	-3	-4	-3	-6	-5	
Tyr	0	-5	-5	-3	-3	-3	-4	-4	
Trp	-8	-7	-6	-2	-6	-5	-7	-7	
Cys	Gly	Pro	Ser	Ala	Thr	Asp	Glu	Asn	Gln

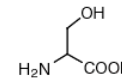
### Small



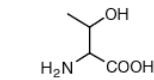
Glycine (Gly, G)  
MW: 57.05



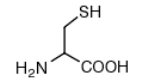
Alanine (Ala, A)  
MW: 71.09



Serine (Ser, S)  
MW: 87.08, pK<sub>a</sub> ~ 16

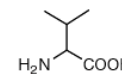


Threonine (Thr, T)  
MW: 101.11, pK<sub>a</sub> ~ 16

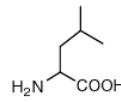


Cysteine (Cys, C)  
MW: 103.15, pK<sub>a</sub> = 8.35

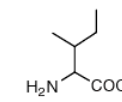
### Hydrophobic



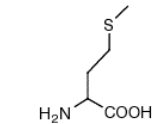
Valine (Val, V)  
MW: 99.14



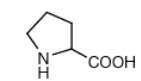
Leucine (Leu, L)  
MW: 113.16



Isoleucine (Ile, I)  
MW: 113.16

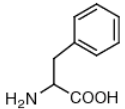


Methionine (Met, M)  
MW: 131.19

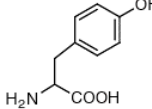


Proline (Pro, P)  
MW: 97.12

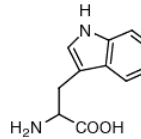
### Aromatic



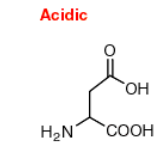
Phenylalanine (Phe, F)  
MW: 147.18



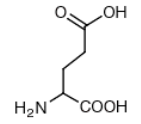
Tyrosine (Tyr, Y)  
MW: 163.18



Tryptophan (Trp, W)  
MW: 186.21

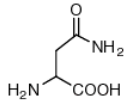


Aspartic Acid (Asp, D)  
MW: 115.09, pK<sub>a</sub> = 3.9

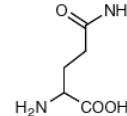


Glutamic Acid (Glu, E)  
MW: 129.12, pK<sub>a</sub> = 4.07

### Amide

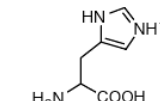


Asparagine (Asn, N)  
MW: 114.11

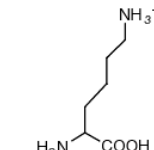


Glutamine (Gln, Q)  
MW: 128.14

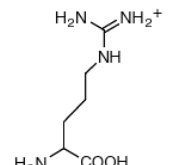
### Basic



Histidine (His, H)  
MW: 137.14, pK<sub>a</sub> = 6.04



Lysine (Lys, K)  
MW: 128.17, pK<sub>a</sub> = 10.79



Arginine (Arg, R)  
MW: 156.19, pK<sub>a</sub> = 12.48

# Verwendung

---

- Welche PAM Matrix soll man 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
  - Iteratives Verfahren einsetzen und auf Konvergenz warten
  - Verschiedene Matrizen testen
  - Externes Wissen (Chemie, Domänen, etc.) hinzuziehen
  - Default nehmen (und beten)

# BLOSUM Matrizen

---

- Henikoff & Henikoff (1992). *Amino acid substitution matrices from protein blocks*. PNAS
- Hauptkritikpunkte am PAM Ansatz
  - PAM-x vervielfältigen Fehler in PAM-1
  - Einbeziehung der kompletten Proteinsequenzen ist zweifelhaft
    - Unterliegen alle Positionen der selben Mutationsrate?
- Neues Verfahren: BLOSUM
  - BLOcks SUbstitution Matrix
  - Erster Schritt: Multiple Alignments evolutionär entfernter, aber homologer Proteinsequenzen
  - Matrixberechnung nur aus konservierten Blöcken
  - Heute populärer als PAM Matrizen

# 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 (später mehr)
  - Heute: Verwendung weiterer Domänen aus PRINTS, PFAM, ...
- Ein **Block** ist ein zusammenhängendes und hochgradig konserviertes Stück eines MSAs

```
FMYMFYVPL_PQ_QVY
FYQQF_VQLYP_MFQV_
FMY_YUVQQP_UMUQ_
```

# BLOSUM Matrizen

---

- Berechnung der BLOSUM Matrizen: Wie PAM-1 Matrix, aber nur die Sequenzen in den Blocks werden betrachtet
- BLOSUM-x Matrizen
  - Bias durch **viele sehr ähnliche Sequenzen**
    - Sequenzdatenbanken sind keine zufälligen Stichproben
  - 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 hängt mit dem evolutionären Abstand der betrachteten Sequenzen zusammen
- BLOSUM-62 heute typischer Default bei Alignments

# Zusammenfassung

---

- Willkommen im Reich der **Heuristiken**
- Substitutionsmatrizen
  - Beobachtung evolutionär entstandener Unterschiede
  - Schätzung des wahren evolutionären Abstands aus der Zahl der Unterschiede
    - Abstand = Zahl der Punktmutationen  $\neq$  Zeiteinheit
  - Implizite Beachtung vielfältiger Faktoren
    - Ladungen, Nachbarschaft, 3D-Struktur, ...
  - Berechnung von Matrizen immer mit Hilfe von Alignments
    - Iterative Strategien



# Literatur

---

- Ewens, W. J. and Grant, G. R. (2001). "Statistical Methods in Bioinformatics", Springer.
- Dayhoff, M. O., R. V. Eck, C. M. Park. 1972. A model of evolutionary change in proteins. In M. O. Dayhoff, ed., Atlas of Protein Sequence and Structure Vol. 5.
- Henikoff, S. and Henikoff, J. G. (1993). "Performance evaluation of amino acid substitution matrices." Proteins 17(1): 49-61.

# Selbsttest

---

- Was will das Jukes-Cantor Modell berechnen?
- Warum spricht man dabei immer vom „Erwartungswert“. Kann man die Zahl der Mutationen nicht exakt bestimmen?
- Was besagt die Molecular Clock Theory?
- Wie werden PAM Matrizen berechnet?
- Wie werden BLOSSUM Matrizen berechnet (und warum reicht PAM nicht)?