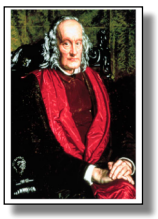
Quantifizierung evolutionärer Veränderungen

- Begriff der Homologie/Homoplasie
- Methoden des Sequenzvergleichs/Alignments
- Verfahren und Modelle zur Berechnung von Austauschraten in DNA und Proteinen

Thomas Hankeln, Institut für Molekulargenetik SS 2010 JOHANNES GUTENBERG UNIVERSITÄT MAINZ

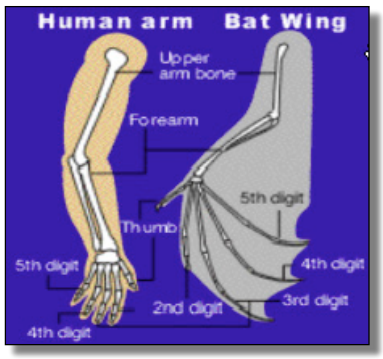
Der Begriff der Homologie



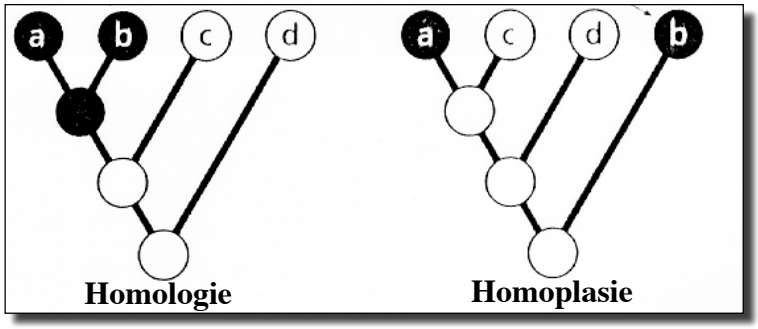
Richard Owen 1843

Homology: **„the same organ under every variety of form and function (true correspondence)“**

Analogy: **„superficial or misleading similarity**



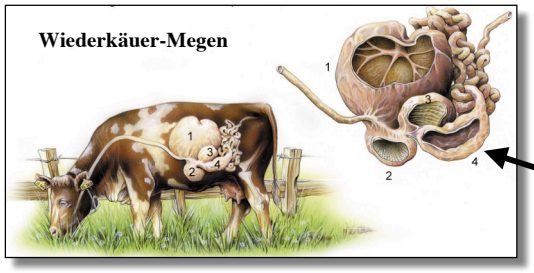
Homologie vs. Homoplasie/Konvergenz



- Merkmal ‚Schwarz' von gemeinsamem Vorläufer geerbt

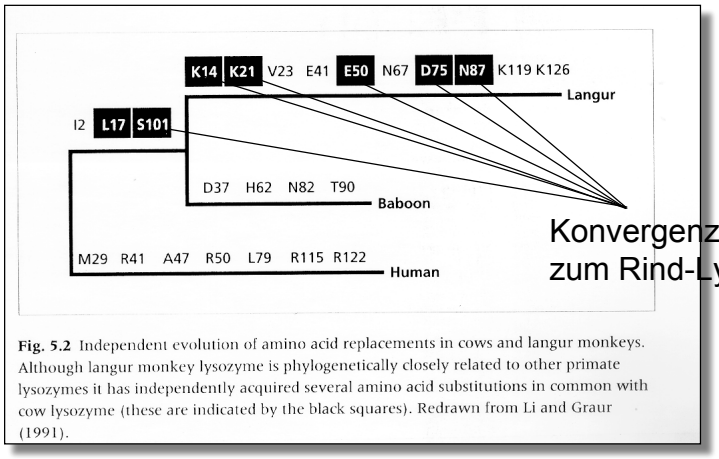
- Merkmal ‚Schwarz' konvergent aus ‚weißen' Vorläufern entstanden

Beispiel: Konvergente Evolution in Proteinen



Spezielles Lysozym zum Verdau von Bakterien, die in bestimmten Magen-Abschnitten für den Aufschluss der pflanzlichen Nahrung sorgen

Beispiel: Konvergente Evolution in Proteinen



Konvergenzen zum Rind-Lysozym

Fig. 5.2 Independent evolution of amino acid replacements in cows and langur monkeys. Although langur monkey lysozyme is phylogenetically closely related to other primate lysozymes it has independently acquired several amino acid substitutions in common with cow lysozyme (these are indicated by the black squares). Redrawn from Li and Graur (1991).

Homologie, Identität, Ähnlichkeit

Beim Vergleich zwischen DNA-Sequenzen oder Proteinsequenzen sprechen wir zunächst immer von

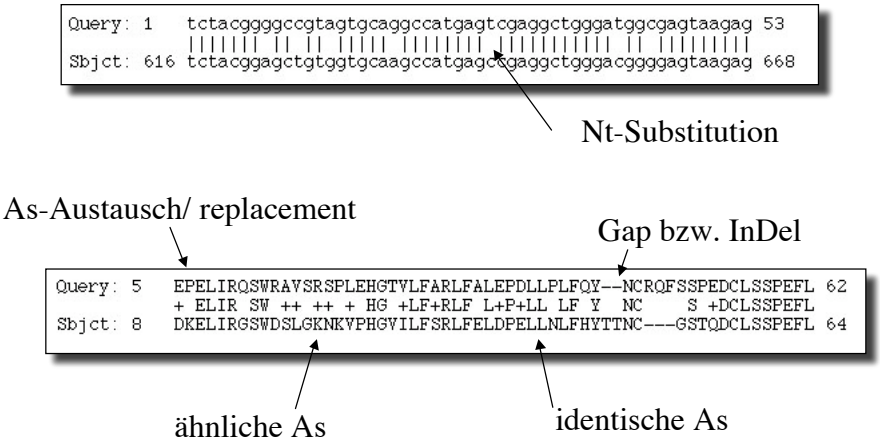
- **Sequenzübereinstimmung (identity) oder**
- **Sequenzähnlichkeit (similarity)**

Erst aus diesem Vergleich heraus können wir überlegen, ob die gefundenen Übereinstimmungen wirklich **homolog** sind!

Wie gehen wir also vor…?

1. Wir erstellen ein Sequenzalignment
2. Wir schließen daraus auf ‚Homologie‘ (bei Gegenteil ist Vergleich sinnlos)
3. Wir berechnen die Evolutionsereignisse, die ‚wirklich stattgefunden haben‘
4. Wir können mit diesen Daten z.B. Stammbäume rekonstruieren oder Evolutionsereignisse datieren…

Vergleich von DNA- oder Proteinsequenzen durch „Alignment“

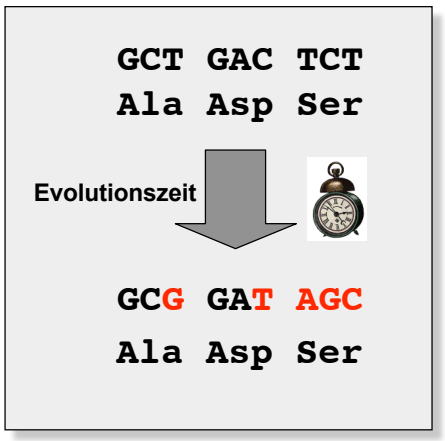


Protein-Sequenzen: Identität & Ähnlichkeit

```
 Score = 91.3 bits (223), Expect = 4e-18
 Identities = 59/156 (37%), Positives = 88/156 (55%), Gaps = 14/156 (8%)

Query: 4   MYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVAGLN 63
           +YKKI+ PTD S+ +  A KH          EV  ++V+D          S  +G+
Sbjct: 25  LYKKIVIPTDGSDVSLEAAKHAINIAKEFDAEVYAIYVVD---------VSPFVGLPA-- 73

Query: 64  KSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEGVDI 123
           +   E  +EL   L EE +  ++ +KK  E+ G K+   ++ G+P  EIV+ AE +  D+
Sbjct: 74  EGSWELISEL---LKEEGQEALKKVKKMAEEWGVKIHTEMLEGVPANEIVEFAEKKKADL 130

Query: 124 IIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKR 159
           I+MG+ GKT L+ ILLGSV E VIK ++ PVLVVK+
Sbjct: 131 IVMGTTGKTGLERILLGSVAERVIKNAHCPVLVVKK 166
```

Bei Proteinsequenz-Alignments unterscheidet man **Sequenzidentität** und **Sequenzähnlichkeit** (= Identität plus iso-funktionelle As)

Wann DNA? Wann Protein?



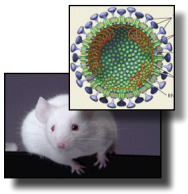
Während der Evolution wird die DNA durch ‚stille' Mutationen stark verändert, während die Selektion die Veränderung auf Aminosäureebene weitgehend verhindert:

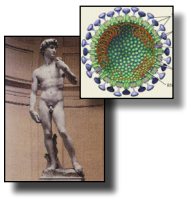
• Vergleich auf DNA-Ebene funktioniert nur zwischen nahe verwandten Taxa/ Genen

• Vergleich auf Aminosäureebene kann noch Ähnlichkeiten von entfernt verwandten Sequenzen detektieren

Wann DNA? Wann Protein?

Eng verwandte SARS-Varianten
in der Population





Corona-Virus-Gruppen
aus verschiedenen Spezies

Alignment = Evolutionshypothese

Wir treffen durch die
Wahl des Alignments
eine Annahme über den
Ablauf der Evolution!!!!

Warum ist ein „richtiges" Alignment so problematisch?

- **Zwei beliebige Sequenzen lassen sich prinzipiell immer alignen!**
- **Es gibt viele mögliche Alignments**
- **Sequenz-Alignments müssen also in ihrer ‚Güte' bewertet werden, um das ‚optimale Alignment' zu finden**
- **Häufig wird es mehrere gleich gute Lösungen geben**

```
ACGTACGTACGTACGTACGTACGT
   |   |   |   |   |   |
 GATCGATCGATCGATCGATCGATC
```

```
ACGTACGTACGTACGTACGTACGT
   |   |   |   |   |   |
 GATCGATCGATCGATCGATCGATC
```

…etwas einfacher geht's mit dem 20 As-Alphabet von Proteinen (manchmal)

Finde das optimale Alignment:

```
THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
THIS IS A SHORT SENTENCE


THIS IS A RATHER LONGER - SENTENCE THAN THE NEXT
|||| || | --*|-- -|---| - |||||||| ---- --- ----
THIS IS A --SH-- -O---R T SENTENCE ---- --- ----
or
THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
|||| || | ------ ------ |||||||| ---- --- ----
THIS IS A SHORT- ------ SENTENCE ---- --- ----
```

Wie erstellt man ein möglichst „richtiges" Alignment ?

Wir brauchen **„evolutionäre Modelle"** (quasi Spielregeln),
um die beobachteten Sequenzveränderungen richtig zu bewerten:

- wie häufig mutiert ein A nach G bzw. nach C oder T (Transitionen : Transversionen)?
- wie häufig entstehen In/Dels relativ zu Substitutionen?
- wie häufig wird während der Proteinevolution z. B. ein Tryptophan durch irgendeine andere Aminosäure ersetzt?

...zunächst zur Behandlung von Lücken!

Ein einfacher Score-Wert zur Bewertung eines Alignments...

$$S = Y - \sum W_k$$

S = Similarity-Score
Y = Anzahl an Matches

W_k = gap penalty für gaps der Länge k

Das Setzen einer Lücke wird durch einen negativen Score (gap penalty) bestraft!

Auswirkungen der gap penalty

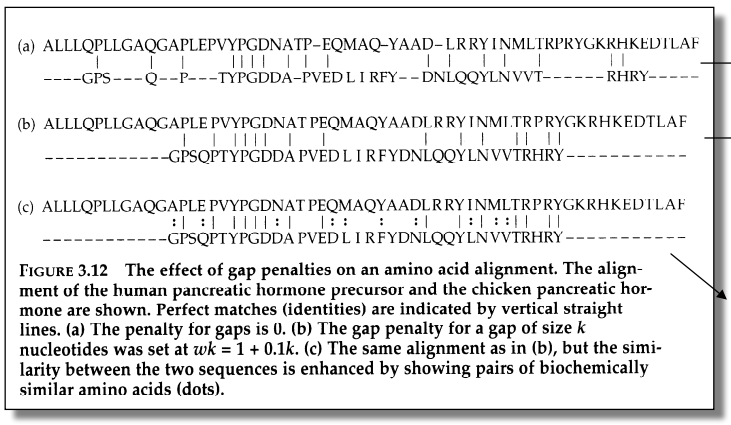


FIGURE 3.12 The effect of gap penalties on an amino acid alignment. The alignment of the human pancreatic hormone precursor and the chicken pancreatic hormone are shown. Perfect matches (identities) are indicated by vertical straight lines. (a) The penalty for gaps is 0. (b) The gap penalty for a gap of size k nucleotides was set at $wk = 1 + 0.1k$. (c) The same alignment as in (b), but the similarity between the two sequences is enhanced by showing pairs of biochemically similar amino acids (dots).

→ Penalty = 0

→ Penalty $w_k=1+0.1k$

Anzeigen der biochemisch verwandten As macht deutlich, daß das Alignment (b) Sinn macht

…und jetzt zu den Austauschen!

• in sog. „Substitutionsmatrizen“ wird die relative Häufigkeit erfasst, mit der Nukleotide oder Aminosäuren während der Evolution ausgetauscht werden.

⇒ Daraus werden „Belohnungswerte“ oder „Kosten“ errechnet, die uns helfen, ein bestmögliches Alignment zu erstellen

Eine einfache Identitätsmatrix bei Nukleotidsequenzen...

	A	C	G	T
A	1			
C	0	1		
G	0	0	1	
T	0	0	0	1

• alle Richtungen von Nt-Austauschen sind gleich wahrscheinlich

• bei jedem „match“ beider Sequenzen gibt es 1 Punkt für den Übereinstimmungs-Score

DNA-Alignment-Bewertung

```
seqA   TCAGACGATTG (11)
seqB   TCGGAGCTG (9)

I.     TCAG-ACG-ATTG
       TC-GGA-GC-T-G

II.    TCAGACGATTG
       TCGGAGCTG--

III.   TCAG-ACGATTG
       TC-GGA--GCTG
```

D = 7 – 6(3+1x0.1) = - 11.6

D = 4 – (3+2x0.1) = **+ 0.8**

D = 6 – 2(3+1x0.1) – (3+2x0.1)
= - 3.4

Match = +1
Gap-Parameter:
d = 3 (gap opening)
e = 0.1 (gap extension)

Bei hoher gap opening penalty!

DNA-Alignment-Bewertung

```
seqA   TCAGACGATTG (11)
seqB   TCGGAGCTG (9)

I.     TCAG-ACG-ATTG
       TC-GGA-GC-T-G

II.    TCAGACGATTG
       TCGGAGCTG--

III.   TCAG-ACGATTG
       TC-GGA--GCTG
```

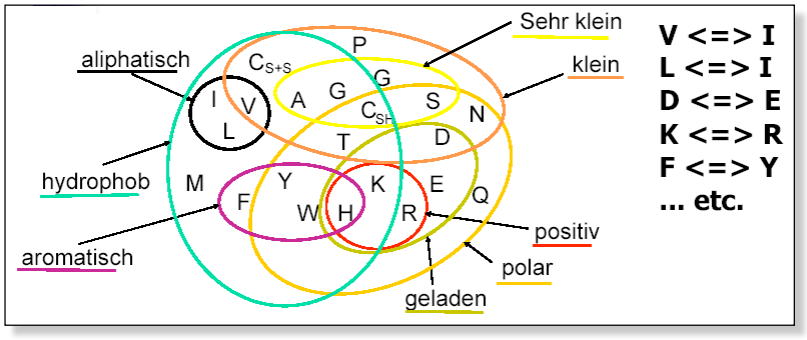
?

Match = +1
Gap-Parameter:
d = 1 (gap opening)
e = 0.1 (gap extension)

Bei niedriger gap opening penalty!

Substitutionsmatrizen bei Proteinen sind komplizierter

- bei **Proteinen** gibt es **20 As**
- chemisch-funktionelle Ähnlichkeit bestimmt Wahrscheinlichkeit eines Austauschs während der Evolution.



V <=> I
L <=> I
D <=> E
K <=> R
F <=> Y
... etc.

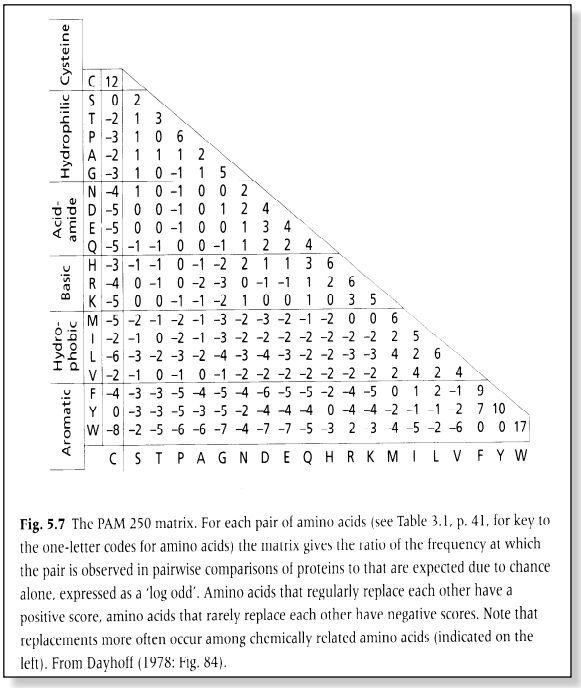
Substitutions-Matrizen für Proteine

- chemisch-funktionelle Ähnlichkeit der As bestimmt Wahrscheinlichkeit eines Austauschs während der Evolution. Daher...

- ...sind die „Kosten“ bzw. die „Belohnung“ für bestimmte Austausche unterschiedlich hoch!

- Definition von Kosten bzw. Belohnung erfolgt über Matrizen:

  > **PAM-Matrizen** (Dayhoff 1978)
  > **BLOSUM-Matrizen** (Henikoff & Henikoff 1992)
  u. einige mehr



		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
Cysteine	C	12																			
Hydrophilic	S	0	2																		
	T	−2	1	3																	
	P	−3	1	0	6																
	A	−2	1	1	1	2															
	G	−3	1	0	−1	1	5														
Acid-amide	N	−4	1	0	−1	0	0	2													
	D	−5	0	0	−1	0	1	2	4												
	E	−5	0	0	−1	0	0	1	3	4											
	Q	−5	−1	−1	0	0	−1	1	2	2	4										
Basic	H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
	R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
	K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
Hydro-phobic	M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
	I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
	L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
	V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
Aromatic	F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
	Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	2	7	10	
	W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	3	4	−5	−2	−6	0	0	17

Fig. 5.7 The PAM 250 matrix. For each pair of amino acids (see Table 3.1, p. 41, for key to the one-letter codes for amino acids) the matrix gives the ratio of the frequency at which the pair is observed in pairwise comparisons of proteins to that are expected due to chance alone, expressed as a 'log odd'. Amino acids that regularly replace each other have a positive score, amino acids that rarely replace each other have negative scores. Note that replacements more often occur among chemically related amino acids (indicated on the left). From Dayhoff (1978: Fig. 84).

PAM - Matrix

- PAM = percent accepted mutation

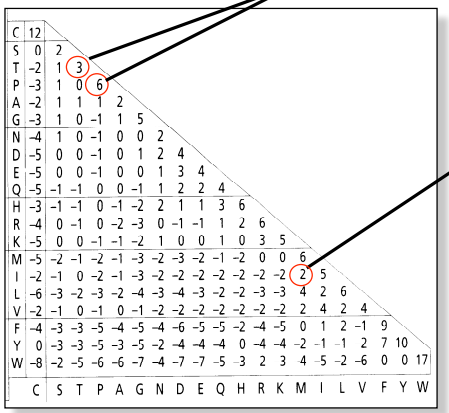
- positiver Wert = hohe Wahrscheinlichkeit, dass die As während der Evolution wegen ähnlicher Funktion erhalten bleiben:

➢sollten also im Alignment gegenüberstehen (ergibt ‚Belohnung')

Bewertung eines As-Alignments

Sequenz 1
Sequenz 2

```
PTHPLASKTQILPEDLASEDLTI
||||||     |     ||  ||
PTHPLAGERAIGLARLAEEDFGM
```



		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
	C	12																			
	S	0	2																		
	T	−2	1	(3)																	
	P	−3	1	0	(6)																
	A	−2	1	1	1	2															
	G	−3	1	0	−1	1	5														
	N	−4	1	0	−1	0	0	2													
	D	−5	0	0	−1	0	1	2	4												
	E	−5	0	0	−1	0	0	1	3	4											
	Q	−5	−1	−1	0	0	−1	1	2	2	4										
	H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
	R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
	K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
	M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
	I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	(2)	5						
	L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
	V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
	F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
	Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	2	7	10	
	W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	3	4	−5	−2	−6	0	0	17

```
P:P = +6
T:T = +3
...
I:M = +2


Score =
6+3+...+2 = XX
```

Verändert nach Folie von Stefan Wiemann, DKFZ

Und noch einmal…

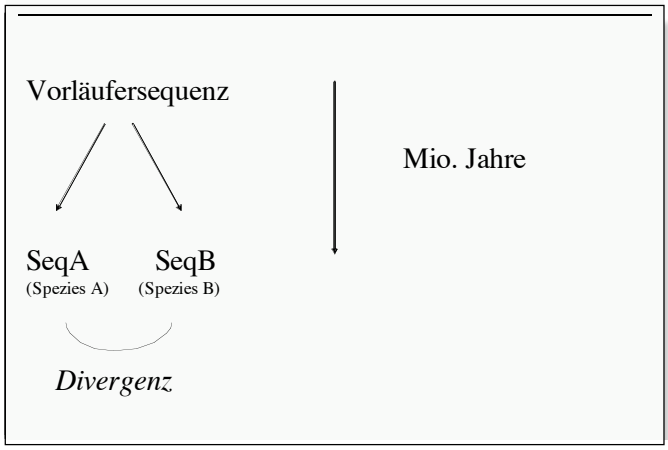
Das korrekte Alignment ist die Basisvoraussetzung für evolutionären Vergleich von Sequenzen!!!

Jedes unserer Alignments ist nur eine evolutionäre Hypothese. Es ist nur so „richtig“, wie die Annahmen richtig sind, die wir zugrundelegen!!!

Wie gehen wir vor…?

1. Wir erstellen ein Sequenzalignment
2. Wir schließen daraus auf ‚Homologie‘ (bei Gegenteil ist Vergleich sinnlos)
3. Wir berechnen die Evolutionsereignisse, die ‚wirklich stattgefunden haben‘
4. Wir können mit diesen Daten z.B. Stammbäume rekonstruieren oder Evolutionsereignisse datieren…

Veränderungen in Sequenzen während der Evolution



Ziel: Berechnen der Evolutonsereignisse, die ‚wirklich statt gefunden haben'

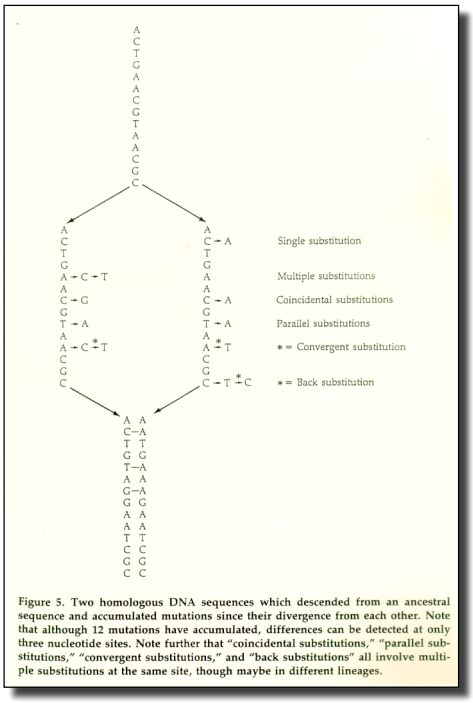


Figure 5. Two homologous DNA sequences which descended from an ancestral sequence and accumulated mutations since their divergence from each other. Note that although 12 mutations have accumulated, differences can be detected at only three nucleotide sites. Note further that "coincidental substitutions," "parallel substitutions," "convergent substitutions," and "back substitutions" all involve multiple substitutions at the same site, though maybe in different lineages.

Multiple Austausche

Problem:
Die sichtbaren Austausche zeigen nicht den wahren Umfang der Ereignisse während der Evolution!

Je mehr Evolutionszeit vergangen ist, desto höher ist die Chance, daß es an bestimmten Positionen multiple Austausche gegeben hat.

Die beobachteten Divergenzwerte müssen hochkorrigiert werden…

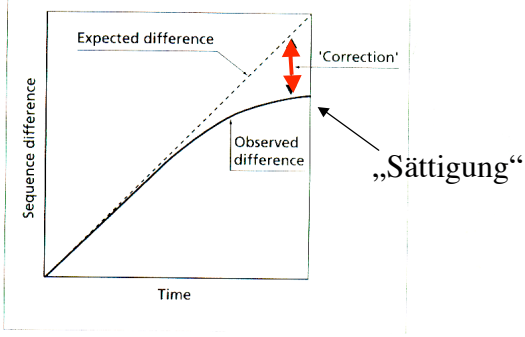


Fig. 5.12 The need to correct observed sequence differences. The extent of observed differences between two sequences is not linear with time (as we would expect if the rate of molecular evolution is approximately constant) but curvilinear due to multiple hits. The goal of distance correction methods is to recover the amount of evolutionary change that the multiple hits have overprinted and to 'correct' the distances for unobserved hits. In effect, the methods seek to 'straighten out' the line representing observed differences.

…dies betrifft besonders Nt-Sequenzen!

Die beobachteten Divergenzwerte müssen hochkorrigiert werden…

Q: Aber wie können wir die „wahren“, stattgefundenen Austausche extrapolieren?

A: Wir müssen die richtigen Annahmen über den Verlauf der Evolution treffen (= wiederum Modelle der Sequenzevolution zugrundelegen).

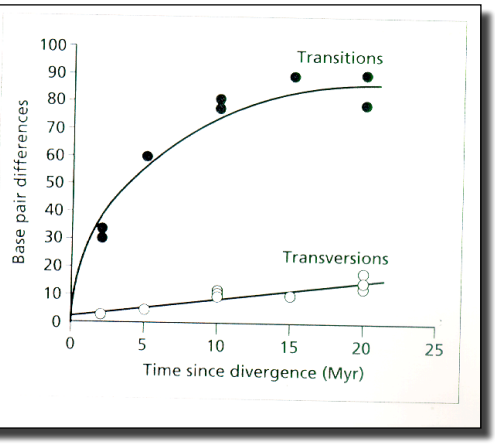
z. B. ein bestimmtes Transitions/Transversions-Verhältnis, oder bestimmte bevorzugte Richtungen von Mutationen (zB GC > AT)

Die Art der Annahmen können wir zuvor aus unseren Sequenz-vergleichsdaten ermitteln.

Unser Evolutionsmodell hier…

mtDNA bei Säugern

Fig. 5.13 The number of transitions and transversions between the same bovid mammal sequences used in Fig. 5.11. Transitions accumulate much more rapidly than transversions and become saturated, whereas transversions accumulate more slowly and show no evidence of saturation.



- Transitionen akkumulieren schnell > Sättigung
- Transversionen akkumulieren langsam und proportional zur Zeit

Modelle für die Evolution von Nukleotidsequenzen

- Jukes-Cantor (1969) one-parameter model (JC)
- Kimura two-parameter (K2P)
- Felsenstein 81
- Hasegawa, Kishino, Yano (HKY85)
- General time-reversible model (REV, GTR)

und viele andere…

http://hcv.lanl.gov/content/hcv-db/findmodel/matrix/all.html
http://workshop.molecularevolution.org/resources/models/dnamodels.php

Modelle für die Evolution von Nukleotidsequenzen

Alle diese Modelle treffen nur Annahmen für Nukleotid-substitutionen!!

Indel-Positionen werden nicht berücksichtigt. Sie werden sogar zumeist aus dem Sequenzvergleich entfernt!

- „complete deletion“ > sinnvoll wenn alignment in Bereichen mit Lücken unsicher ist
- „pairwise deletion“ > bei kleinen gaps, die statistisch über das Alignment verteilt sind

Modelle für die Evolution von Nukleotidsequenzen

- wie ist die Wahrscheinlichkeit, daß ein Nukleotid i zum Nukleotid j wird?

Parameter:
> Ausgangsfrequenz der einzelnen Basen
> Transitions/Transversions-Verhältnis
> individuelle Mutabilität jedes Nukleotids in jedes andere

Die Modelle treffen je nach Kompliziertheits-Grad für wenige oder viele dieser Parameter eine Annahme.

Das Jukes-Cantor (JC) „one parameter"-Modell

- alle 4 Basen haben dieselbe Frequenz
- alle Substitutionen sind gleich wahrscheinlich

$$K = - 3/4 \ln (1 - 4/3 p)$$

K = subst./ site
p = diverg. Posit./ Gesamtzahl der Nukleotide
(unkorrigierte „Hamming"-Distanz)

Achtung: bei p > 3/4 wird Formel ungültig!!!

Das Kimura „two-parameter"-Modell (K2P)

- alle 4 Basen haben dieselbe Frequenz
- Transitionen und Transversionen haben unterschiedliche Häufigkeiten

$$K = 1/2 \ln (1 / (1 - 2P - Q)) + 1/4 \ln (1 / (1 - 2Q))$$

P = divergente Ti pro Nt_{gesamt}
Q = div. Tv pro Nt_{gesamt}

JC vs. K2P

- Beispiel 1 : 2 Seq mit je 200 Bp
  Divergenz 20 Ti, 4 Tv

  P (unkorrigiert) = 24 / 200 = 0,12
  K (JC) = 0,13
  K (K2P) = 0,13

Wenn p <<1 (Divergenz sehr klein) ist, kann man das einfachste Modell (JC) nehmen oder gar auf Korrektur verzichten

JC vs. K2P

- Beispiel 2 : 2 Seq mit je 200 Bp
  Divergenz 50 Ti, 16 Tv

  P (unkorrigiert) = 66 / 200 = 0,33
  K (JC) = 0,43
  K (K2P) = 0,48

Wenn Divergenz > 0,15 ist und ein unterschiedliches Ti/Tv-Verhältnis vorliegt, sollte man mindestens K2P-Korrektur durchführen

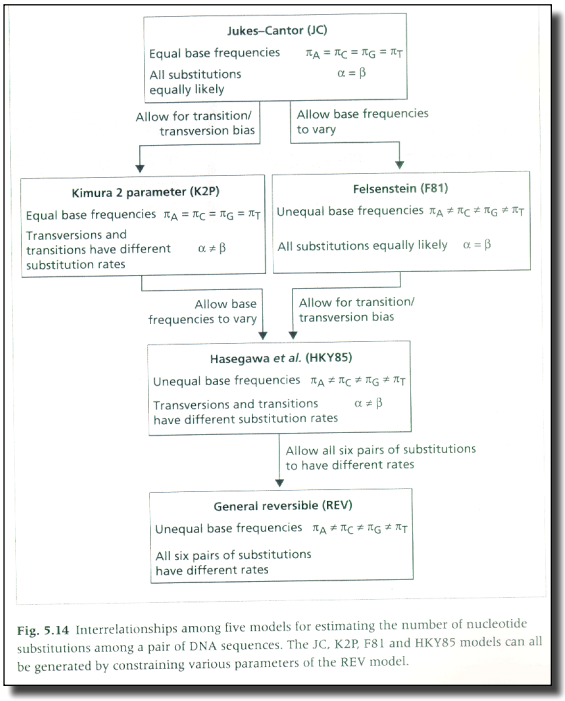


Fig. 5.14 Interrelationships among five models for estimating the number of nucleotide substitutions among a pair of DNA sequences. The JC, K2P, F81 and HKY85 models can all be generated by constraining various parameters of the REV model.

superModels...

Die komplizierteren Modelle treffen noch mehr Annahmen und leiten sich aus den einfachen Modellen ab

Welches Modell soll man denn anwenden?

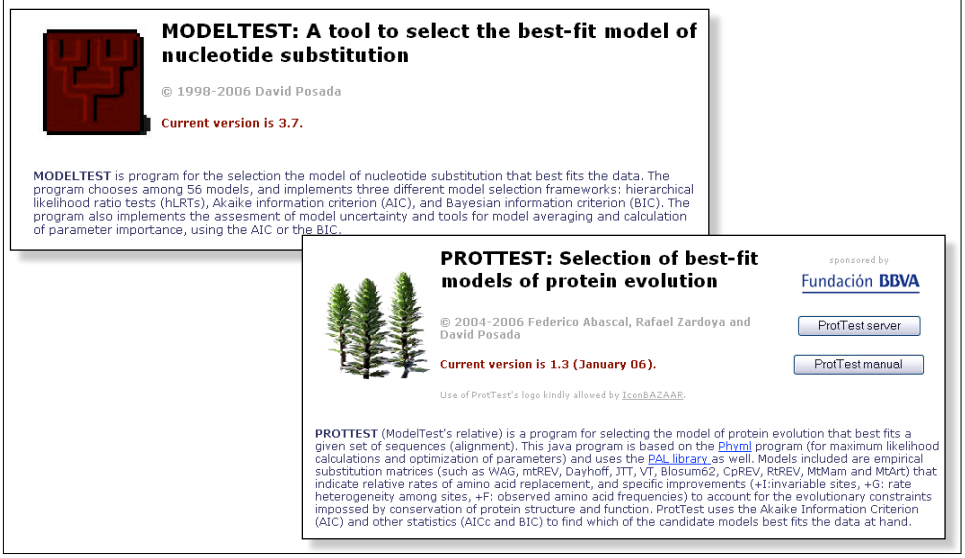
• je mehr Annahmen man trifft (komplizierte Modelle), desto exakter und realistischer sollte unsere Berechnung der Substitutionsrate ausfallen

ABER:

• Zusätzliche Parameter für komplizierte Modelle müssen wir aus unseren Daten abschätzen. Je mehr Annahmen man trifft (und Parameter einbezieht), desto größer wird der statistische Fehler (Varianz) unserer erhaltenen Werte!

Also: die niedrigste Zahl von Parametern (= das einfachste Modell“) nehmen, ohne jedoch die Exaktheit zu verlieren.

Welches Modell soll man denn anwenden?



Austauschberechnungen in proteinkodierenden Genen

… erfordern spezielle Methoden, da synonyme und nicht-synonyme Kodonpositionen nach unterschiedlichen Gesetzmäßigkeiten evolvieren!

Austauschberechnungen in proteinkodierenden Sequenzen

- separat für syn und non-syn Austausche
- ATG & STOP-Kodons ausschließen, da invariabel
- bei mehreren Austauschen in 2 verglichenen Kodons müssen verschiedene Pfade der Evolution getrennt kalkuliert werden:

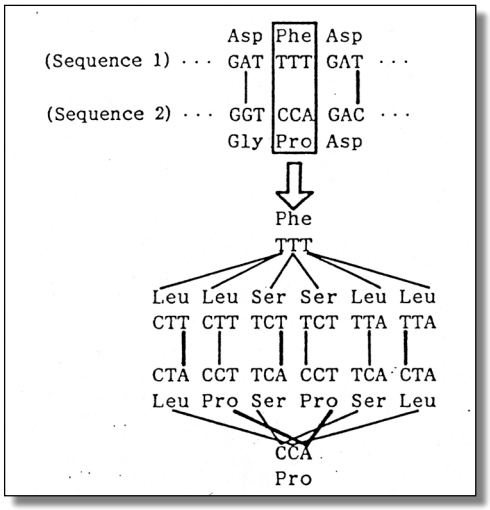
Bsp.1	GTC Val	>	GTT Val	(1 syn)
Bsp. 2	AAT Asn	>	ACG Thr	?

Austauschberechnungen in proteinkodierenden Sequenzen

- bei mehreren Austauschen in 2 verglichenen Kodons müssen verschiedene Pfade der Evolution getrennt kalkuliert werden:

Bsp. 2	AAT Asn		>	ACG Thr		?
Pfad I	AAT Asn	nonsyn >		ACT Thr	syn >	ACG Thr
Pfad II	AAT Asn	nonsyn >		AAG Lys	nonsyn >	ACG Thr

Austauschberechnungen in proteinkodierenden Sequenzen



- bei drei Austauschen in den verglichenen Kodons gibt es sogar 6 mögliche Pfade der Evolution

Austauschberechnungen in proteinkodierenden Sequenzen

- verschiedene Pfade sind unterschiedlich wahrscheinlich
- zwei Berechnungsmethoden zur Auswahl:
  1. „unweighted pathway methods“
  2. „weighted pathway methods“

Austauschberechnungen in proteinkodierenden Sequenzen:
„N-G unweighted pathway method"

Nei & Gojobori MBE 3 (1986) pp.418

Hiv-web.lanl.gov/SNAP/WEBSNAP/SNAP.html

Schritt 1: Ermittlung der „potentiell syn"- und „potentiell nonsyn"-Positionen der einzelnen Kodons der Sequenzen A und B

	Phe T T T	Leu C T A	Leu T T A
pot. syn	0/3 0/3 1/3	1/3 0/3 3/3	1/3 0/3 1/3
Pot. nonsyn	3/3 3/3 2/3	2/3 3/3 0/3	2/3 3/3 2/3

Σ pot. syn sites $_A$ Σ pot. syn sites $_B$

Σ pot. Nonsyn sites $_A$ Σ pot nonsyn sites $_B$

Austauschberechnungen in proteinkodierenden Sequenzen:
„N-G unweighted pathway method"

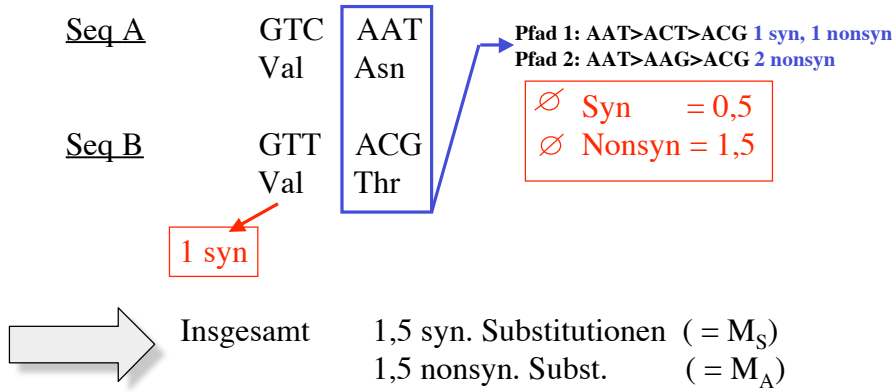
Schritt 2: Berechnung der durchschnittlichen Anzahl an „potentiell syn"- und „potentiell nonsyn"-Positionen der Sequenzen A und B

$\varnothing$ Syn sites $_{A,B}$ $= N_S$

$\varnothing$ Nonsyn sites $_{A,B}$ $= N_A$

Austauschberechnungen in proteinkodierenden Sequenzen:
„N-G unweighted pathway method"

Schritt 3: Klassifizierung der Austausche zwischen A und B



Insgesamt 1,5 syn. Substitutionen (= M_S)
1,5 nonsyn. Subst. (= M_A)

Austauschberechnungen in proteinkodierenden Sequenzen:
„N-G unweighted pathway method"

Schritt 4: Berechnung zunächst der unkorrigierten, dann der für multiple Austausche korrigierten Distanzwerte

➢Unkorrigiert: syn. Subst. / pot. syn. sites $P_S = M_S / N_S$

nonsyn Sub./ pot nonsyn sites $P_A = M_A / N_A$

➢Korrigiert z. B. nach J-C:

$K_S = -3/4 \ln (1 - 4/3 P_S)$

$K_A = -3/4 \ln (1 - 4/3 P_A)$