

Raten und Muster der Evolution von DNA- und Protein-Sequenzen

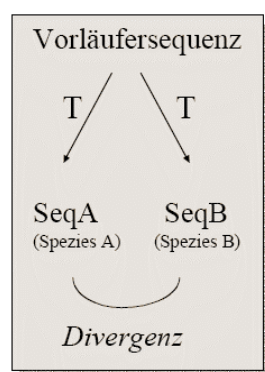
- Substitutionsraten und ihre Variabilität zwischen Genen und innerhalb von Genen
- Gründe für die Variabilität von Substitutionsraten
- Das Konzept der Molekularen Uhr

Thomas Hankeln, Institut für Molekulargenetik

SS 2010

JOHANNES
GUTENBERG
UNIVERSITÄT
MAINZ

Evolutionraten

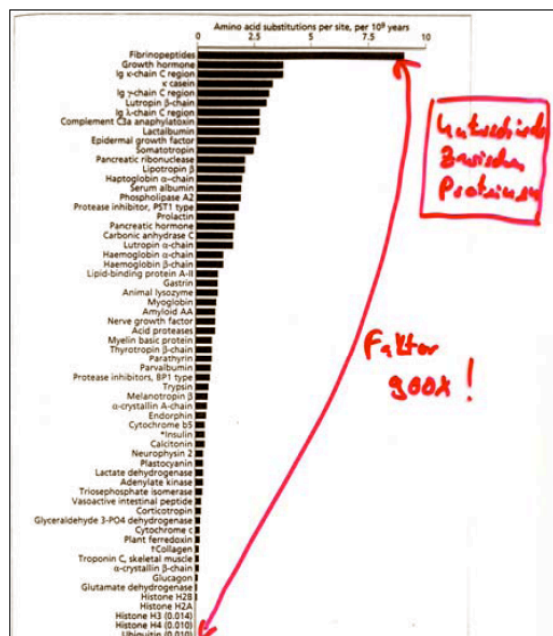


$$\text{Rate } R = K / 2 T$$

Bei Nukleotidsequenzen:

$K_A = d_N$ nicht-syn. Substitutionsrate
 $K_S = d_S$ syn. Substitutionsrate

Erst einmal die Daten....



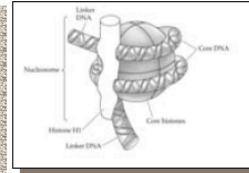
Evolutions- raten in Protein- kodierenden Genen

...beträchtliche
Unterschiede!!

Fig. 7.4 Rates of amino acid substitution for various mammalian proteins. *Excluding guinea pig and corpu, the phylogenetic position of which has been debated (see section 7.4); † Excluding non-repetitive ends. Data from Nei (1987).

Wichtige Proteine evolvieren langsamer!

Protein	Rate (number of aa replacements per site per 10 ⁹ years)
Fibrinopeptide	8.3
Insulin C-peptide	2.4
Ribonuclease	2.1
Hemoglobin	1.0
Cytochrome c	0.3
Histone H4	0.01



courtesy of Dan Graur

Synonyme und nicht-synonyme Evolutionsraten in Genen

Table 1. Rates of synonymous and nonsynonymous substitutions in various mammalian protein-coding genes.*

Gene	L ^b	Nonsynonymous rate (x 10 ³)	Synonymous rate (x 10 ³)
HISTONES			
Histone 3	135	0.00 ± 0.00	6.38 ± 1.19
Histone 4	101	0.00 ± 0.00	6.12 ± 1.32
CONTRACTILE SYSTEM PROTEINS			
Actin α	376	0.01 ± 0.01	3.68 ± 0.43
Actin β	349	0.03 ± 0.02	3.13 ± 0.39
HORMONES, NEUROPEPTIDES, AND OTHER ACTIVE PEPTIDES			
Somatostatin-28	28	0.00 ± 0.00	3.97 ± 2.66
Insulin	51	0.13 ± 0.13	4.02 ± 2.29
Thyrotropin	118	0.33 ± 0.08	4.66 ± 1.12
Insulin-like growth factor II	179	0.52 ± 0.09	2.32 ± 0.40
Erythropoietin	191	0.72 ± 0.11	4.34 ± 0.65
Insulin C-peptide	35	0.91 ± 0.30	6.77 ± 3.49
Parathyroid hormone	90	0.94 ± 0.18	4.18 ± 0.98
Luteinizing hormone	141	1.02 ± 0.16	3.29 ± 0.60
Growth hormone	189	1.23 ± 0.15	4.95 ± 0.77
Urokinase-plasminogen activator	435	1.28 ± 0.10	3.92 ± 0.44
Interleukin I	265	1.42 ± 0.14	4.60 ± 0.65
Relaxin	54	2.51 ± 0.37	7.49 ± 6.10
HEMOGLOBINS AND MYOGLOBIN			
α-globin	141	0.55 ± 0.11	5.14 ± 0.90
Myoglobin	153	0.56 ± 0.10	4.44 ± 0.82
β-globin	144	0.80 ± 0.13	3.05 ± 0.56

(Continued on next page)

Table 1. (Continued)

Gene	L ^b	Nonsynonymous rate (x 10 ³)	Synonymous rate (x 10 ³)
APOLIPROTEINS			
E	283	0.98 ± 0.10	4.04 ± 0.53
A-I	243	1.57 ± 0.16	4.47 ± 0.66
A-IV	371	1.58 ± 0.12	4.15 ± 0.47
IMMUNOGLOBULINS			
Ig V _H	100	1.07 ± 0.19	5.66 ± 1.36
Ig γ1	321	1.46 ± 0.13	5.11 ± 0.64
Ig κ	106	1.87 ± 0.26	5.90 ± 1.27
INTERFERONS			
α1	166	1.41 ± 0.13	3.53 ± 0.61
β1	159	2.21 ± 0.24	5.88 ± 1.08
γ	136	2.79 ± 0.31	8.59 ± 2.56
OTHER PROTEINS			
Aldolase A	363	0.07 ± 0.03	3.59 ± 0.51
Hydroxanthine phosphoribosyltransferase	217	0.13 ± 0.04	2.13 ± 0.30
Creatine kinase M	380	0.15 ± 0.03	3.08 ± 0.37
Glyceradehyde-3-phosphate dehydrogenase	331	0.20 ± 0.05	2.84 ± 0.37
Lactate dehydrogenase A	331	0.20 ± 0.04	5.03 ± 0.61
Acetylcholine receptor γ subunit	540	0.29 ± 0.04	3.23 ± 0.31
Fibrinogen γ	411	0.55 ± 0.06	5.82 ± 0.67
Albumin	590	0.91 ± 0.07	6.63 ± 0.61
Average ^c		0.85 (0.73)	4.61 (1.44)

* All rates are based on comparisons between human and rodent genes and the time of divergence was set at 80 million years ago. Rates are in units of substitutions per site per 10⁶ years.

^b L = number of codons compared.

^c Average is the arithmetic mean, and values in parentheses are the standard deviations, computed over all genes.

Variation Ratios 100X
Variation Ratios 4X

Die molekulare Evolution ist sehr langsam!

Änderungen auf Nukleotidebene erfolgen mit einer Häufigkeit im Bereich von nur 10^{-9} pro Jahr!

Die molekulare Evolution ist sehr langsam, aber...

...die synonyme Nt-Substitutionsrate ist um ein Mehrfaches höher als die nicht-synonyme Nt-Substitutionsrate:

K_A/d_N	Säuger	$0,85 \pm 0,73 \times 10^{-9}$ Subst/Ort x Jahr
K_S/d_S	Säuger	$4,61 \pm 1,44 \times 10^{-9}$ Subst/Ort x Jahr

Ein typisches Muster...

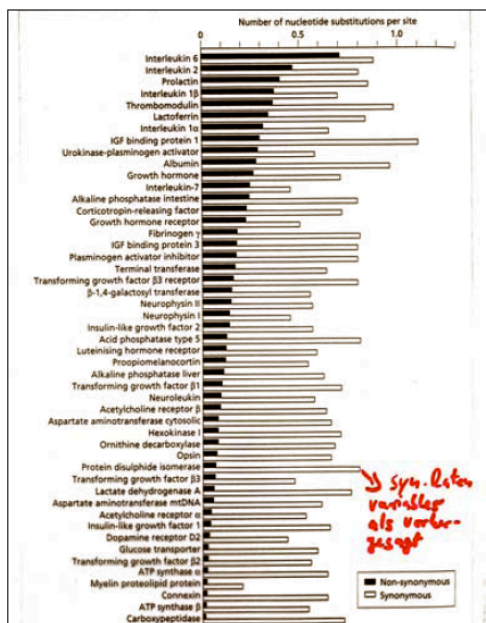
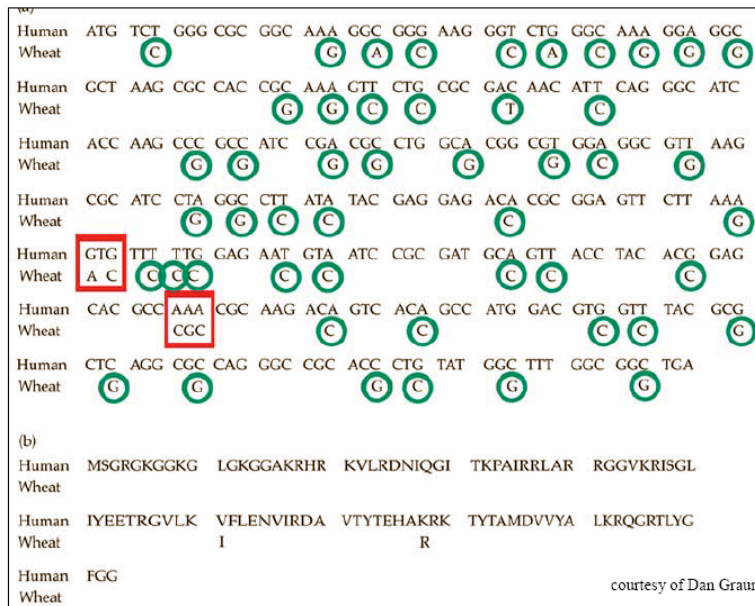


Fig. 7.7 Estimated numbers of non-synonymous (d_n) and synonymous (d_s) substitutions per site for 49 genes from primates, rodents and artiodactyls combined. The data are organized according to the number of non-synonymous changes (highest at top, lowest at bottom). Compare this to the rates of amino acid change shown in Fig. 7.4. Data from Ohno (1995).

Die synonyme
Evolutionsrate
ist zwischen
Genen
erstaunlich
variabel

Warum eigentlich?
(kommt später)

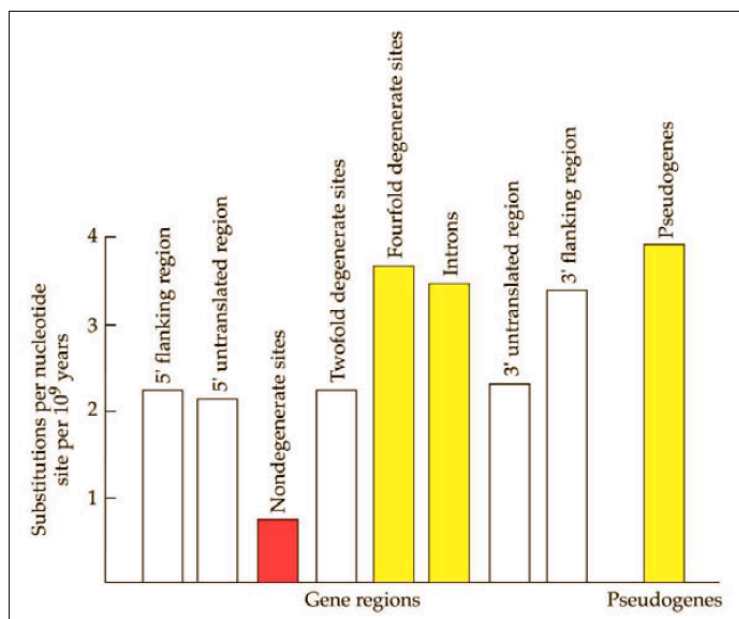
Evolutionraten in nicht-kodierenden Genombereichen

Rates of nucleotide substitution in 5' and 3' untranslated regions and at fourfold degenerate sites of protein-coding genes, based on comparisons between human and mouse or rat genes.^a

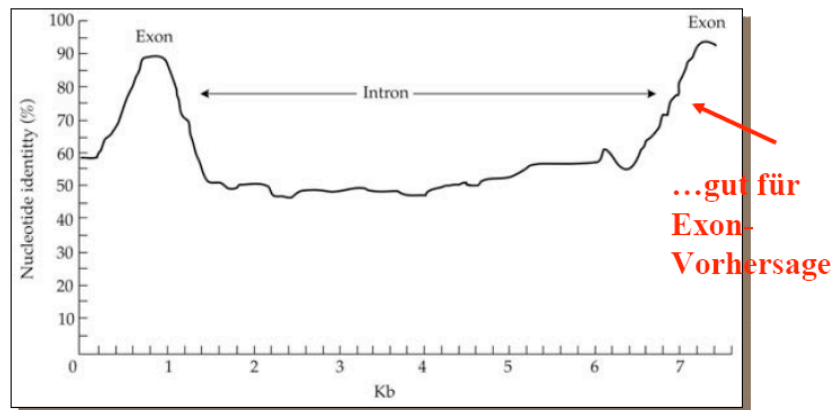
Gene	5' untranslated		3' untranslated		Fourfold degenerate	
	L ^b	Rate	L	Rate	L	Rate
ACTH	99	1.87 ± 0.41	97	2.32 ± 0.49	275	2.78 ± 0.34
Aldolase A	124	1.08 ± 0.26	154	1.73 ± 0.32	195	3.16 ± 0.48
Apolipoprotein A-IV	83	3.06 ± 0.68	134	1.73 ± 0.33	160	3.38 ± 0.50
Apolipoprotein E	23	1.27 ± 0.69	84	1.70 ± 0.42	153	4.00 ± 0.60
Na,K-ATPase β	118	2.45 ± 0.45	1,117	0.57 ± 0.06	118	2.87 ± 0.54
Creatine kinase M	70	1.71 ± 0.46	168	1.79 ± 0.30	178	2.81 ± 0.41
α-fetoprotein	47	3.64 ± 1.13	144	2.79 ± 0.49	225	4.14 ± 0.54
α-globin	34	1.56 ± 0.65	90	2.21 ± 0.50	81	4.47 ± 0.98
β-globin	50	1.30 ± 0.46	126	2.85 ± 0.49	78	2.42 ± 0.56
Glyceraldehyde-3 phosphate dehydrogenase	70	1.34 ± 0.38	121	1.74 ± 0.36	170	2.43 ± 0.39
Growth hormone	21	1.79 ± 0.85	91	1.83 ± 0.41	83	3.82 ± 0.78
Insulin	56	2.92 ± 0.80	53	3.09 ± 0.81	62	4.19 ± 1.00
Interleukin I	59	1.09 ± 0.38	1,046	2.02 ± 0.14	105	2.97 ± 0.60
Lactate dehydrogenase A	95	2.79 ± 0.55	470	2.48 ± 0.23	152	3.64 ± 0.60
Metallothionein II	61	1.88 ± 0.52	111	2.57 ± 0.48	23	2.37 ± 1.00
Parathyroid hormone	84	1.79 ± 0.43	228	2.21 ± 0.30	38	3.85 ± 1.21
Average ^c		1.96 (0.78)		2.10 (0.61)		3.33 (0.69)

^a Rates are in units of substitutions per site per 10⁹ years.
^b L = number of sites.
^c Average is the arithmetic mean, and values in parentheses are the standard deviations, computed over all genes.

Evolutionraten (Überblick)



Kodierende Regionen evolvieren langsamer als nicht-kodierende



z. B. Exons vs. Introns innerhalb eines Gens

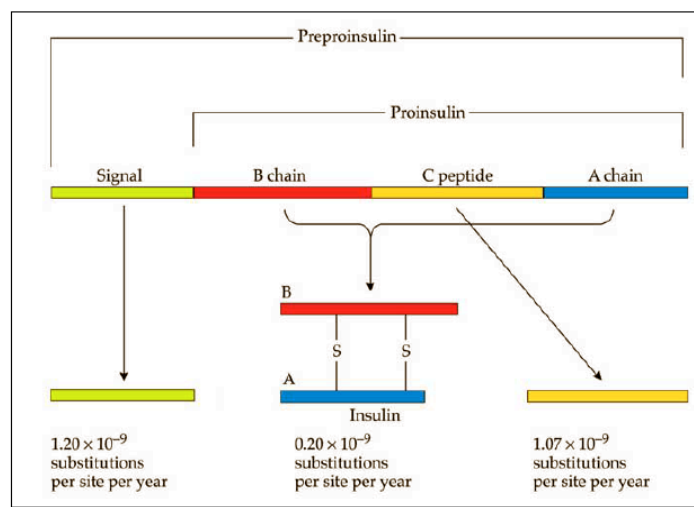
Kann man das auf Regionen
innerhalb von Proteinen
übertragen und so funktionell
wichtige Domänen erkennen?

Beispiel: Prä-Proinsulin

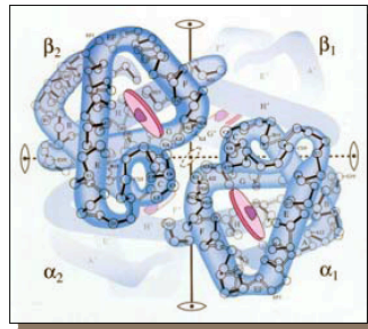
<i>Xenopus</i>	MALWMQCLP-LVLVLLFSTPNTAIAANQHL
<i>Bos</i>	MALWTRLRPLLALLALWPPPPARAFVNQHL
	***** : * * . * : * : . . * : . * : *****
<i>Xenopus</i>	CGSHLVEALYLVCGRGFFYYPKIKRDIEQ
<i>Bos</i>	CGSHLVEALYLVCGERGFFYTPKARREVEG
	***** : ***** * * : * : *
<i>Xenopus</i>	AQVNGPQDNELDG-MQFQPQEYQKMKRGIV
<i>Bos</i>	PQVG---ALELAGGPGAGGLEGPQKRGIV
	. * * . * * * * *
<i>Xenopus</i>	EQCCHSTCSLFQLENYCN
<i>Bos</i>	EQCCASVCSLYQLENYCN
	***** * . * * * : *****

courtesy of Dan Graur

Funktionelle Proteinregionen evolvieren langsamer als funktionell weniger wichtige



Funktionelle Proteinregionen evolvieren langsamer als funktionell weniger wichtige



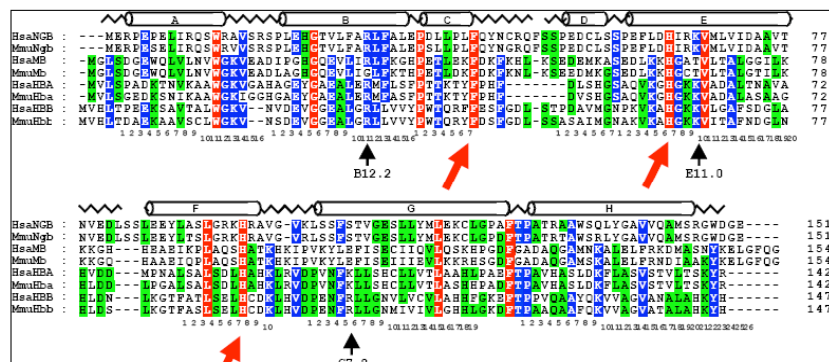
Globine:

- außen: hydrophil
- innen: Hämbindung
Ligandenbindung

As-Substitutions-
Raten:

	surface	heme pocket
α -globin	1.35	0.17
β -globin	2.73	0.24

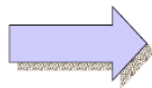
Globinevolution: Korrelation zwischen Austauschrate und Stärke der Selektion



Funktionell sehr
wichtige As-Positionen
sind über mehrere 100
Mio. Jahre konserviert

Der „Selektionismus“ sagt voraus...

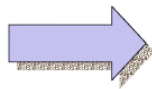
- Mutationen sind entweder vorteilhaft oder nachteilig
- Wenn der Anteil vorteilhafter Mutationen hoch ist, so ist auch die Evolutionsrate hoch
- eine Mutation an einem funktionell wichtigen Ort ist mit größerer Wahrscheinlichkeit vorteilhaft, als eine Mutation an einer nicht-funktionellen Position



Wichtige Abschnitte sollten **schneller** evolvieren als unwichtige!

Die „neutral theory“ sagt voraus...

- Mutationen sind meist entweder nachteilig oder neutral
- Wenn der Anteil nachteiliger Mutationen hoch ist, so ist die Evolutionsrate niedrig
- eine Mutation an einem funktionell wichtigen Ort ist mit größerer Wahrscheinlichkeit nachteilig, als eine Mutation an einer nicht-funktionellen Position



Wichtige Abschnitte sollten **langsamer** evolvieren als unwichtige!

Der selektive Druck auf ein Gen/Protein kann sich während der Evolution ändern

Wie kann man das messen?

- Änderungen der Selektionsbedingungen kann man durch das Verhältnis von nichtsynonymer Rate (d_n bzw. K_a) zu synonymen Substitutionsrate (d_s bzw. K_s) messen
- $d_n \ll d_s$ = **reinigende Selektion**, Konservierung
- $d_n = d_s$ = **neutrale Evolution** beider Kategorien
- $d_n > d_s$ = **positive Selektion**

Lit: L. D. Hurst, Trends in Genetics (2002) 3, S486-487

Bei HIV steigt d_n/d_s nach Transfer über Speziesgrenzen hinweg an

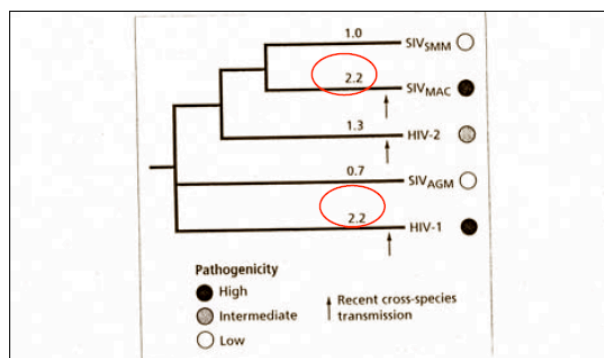
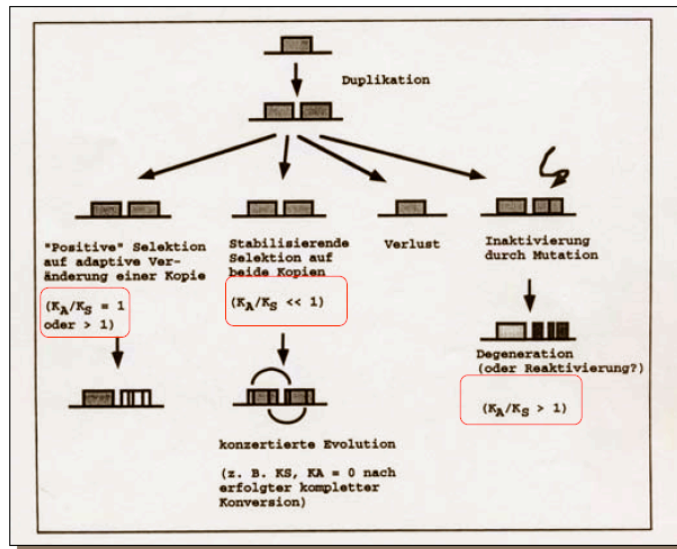
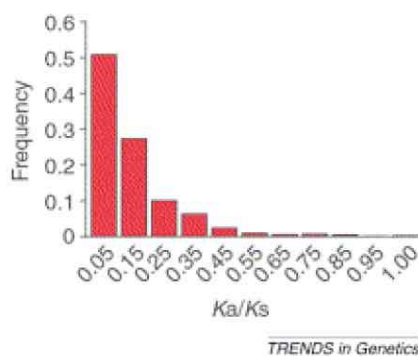


Fig. 7.8 Ratio of non-synonymous (d_n) to synonymous substitutions (d_s) for the envelope protein gp120 from different primate immunodeficiency viruses. The phylogeny presented is a simplified interpretation of the evolutionary history of these viruses (see Chapter 8). HIV-1 and HIV-2 are from humans, SIV_{AGM} from African green monkeys, SIV_{MAC} from macaques and SIV_{SMM} from sooty mangabeys. The higher the ratio, the higher the relative rate of non-synonymous substitution and a ratio > 1.0 indicates the action of positive natural selection (see Box 7.1). Data from Shpaer and Mullins (1993).

dn/ds und das Schicksal von Genduplikaten



Wie häufig ist „positive Selektion“?



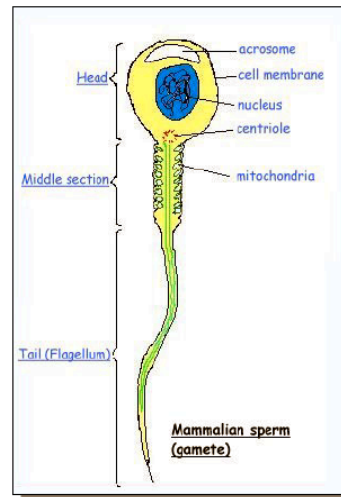
The frequency of different values of K_A/K_S for 835 mouse/rat orthologous genes

Die molekulare Evolution ist größtenteils neutral bzw. negativ selektioniert

Gene unter „positiver Selektion“: Akrosom-Proteine in Spermien

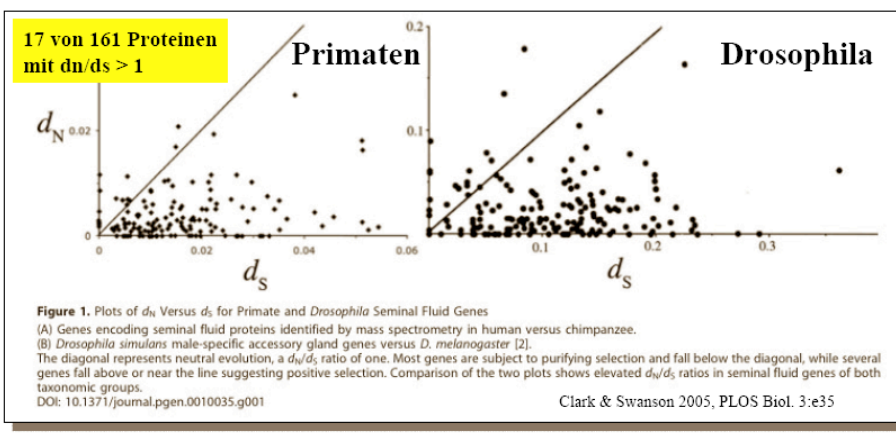
- $dn/ds = 5$ in Genen für
Akrosom-Proteine von Muschel-
Spermien

➤ Variabilität geht mit Speziesbildung
einher, da Änderungen in Akrosom-
Proteinen zur präzygoten
reproduktiven Isolation führen
können



Gene unter „positiver Selektion“: Proteine der Samenflüssigkeit

- Spermien-Kompetition
- Infektionsschutz

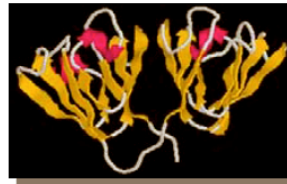


Funktionsänderung erhöht die Evolutionsrate

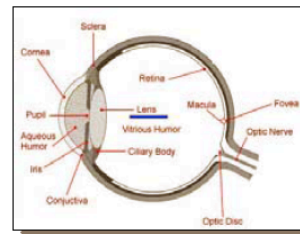


Spalax ehrenbergi, Blindmulle

- unterirdisch lebend
- blind seit > 25 Mio. Jahren



α A-Crystallin



Funktionsänderung erhöht die Evolutionsrate

**α A-Crystallin in Spalax evolviert
20x schneller als in allen anderen
Nagetieren**

Table 1. Number of substitutions per site in α A-crystallin coding regions

	Mouse	Rat	Hamster	Mole rat	Chicken
Mouse		0.000	0.000	0.023	0.084
Rat	0.227		0.000	0.022	0.095
Hamster	0.293	0.344		0.023	0.085
Mole rat	0.534	0.403	0.410		0.112
Chicken	1.407	1.327	1.160	0.983	

K_A^s (upper right part) and K_S^s (lower left part) values are the numbers of nonsynonymous and synonymous substitutions per site, respectively, in the α A coding regions, corrected for multiple events (23, 24).

- Gen-Leserahmen, Spleiß-Signale und Promoter sind jedoch intakt > KEIN Pseudogen!!
- α -Crystallin hat offenbar andere Funktion übernommen (Chaperon; bindet denaturierte Proteine und verhindert deren Aggregation)

Aber Achtung:

Auch ein Funktionsverlust hin zum Pseudogen erhöht die nicht-synonyme Evolutionsrate!!



...zurück ins Labor! Funktionsänderung oder gar -verbesserung biochemisch und funktionell beweisen!

Positive Selektion bei Mensch vs. Chimpanse

Nielsen et al. 2005, PLOS Biol. 3:e170

• 733 von 8079 Genen mit $dn/ds > 1$

Immunabwehr

Geruchsrezeptoren

Spermatogenese

Tumor-Suppression

Apoptose

Gene mit maximaler Expression im Gehirn sind kaum dabei!

Positive Selektion bei Mensch vs. Chimpanse: Ein „genomic conflict“?

Warum Spermatogenese, Tumor-Suppression und Apoptose???

Vorteilhaft für Fortpflanzung:

- weniger Apoptose
- mehr Zellteilung
- = mehr Spermien



Nachteilig für Adulte:

- mehr Krebs

Bedingt die Selektion für Apoptose-Vermeidung
in der Keimbahn (und bessere Fortpflanzung) die hohe
Krebs-Inzidenz beim Menschen??

Nielsen et al. 2005, PLOS Biol. 3:e170

Zusammenfassung

- Die molekulare Evolution ist generell langsam
- Wichtige Gene/Proteine evolvieren langsamer!
(reinigende Selektion)
- Funktionelle Regionen evolvieren langsamer als
nicht-funktionelle *Kimura's First Law of
Molecular Evolution*
- Funktionsverlust (keine Selektion mehr) und
Funktionsänderung (positive Selektion) erhöhen
die Evolutionsrate

Gründe für die Variabilität von Substitutionsraten

Substitutionsrate \approx Mutationsrate
 \approx Probab. Fixierung (+s, o, -s)

A. Variation zwischen Abschnitten eines Gens

- Mutationsrate \pm gleich, d.h. **Selektion** entscheidend (funktionelle Konservierung > langsame Evolution)
- $dn/ds > 1$ ist Hinweis auf “positive Selektion”

Gründe für die Variabilität von Substitutionsraten

Substitutionsrate \approx Mutationsrate
 \approx Probab. Fixierung (+s, o, -s)

B. Variation zwischen Genen

- **Selektion** verantwortlich bei nicht-synonymen Substitutionen

Frage: Warum variiert denn eigentlich die synonyme Substitutionsrate?

Die „neutral theory“ sagt voraus...

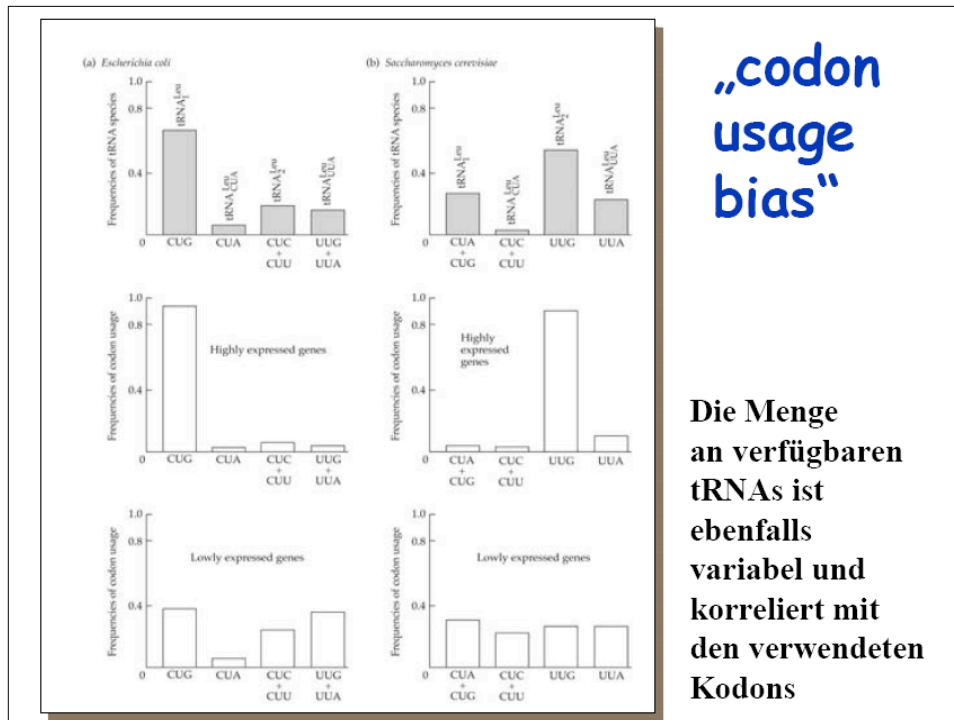
...die Basenzusammensetzung und der Gebrauch alternativer Kodons in Genen wird primär durch **Mutationsprozesse, nicht durch Selektion** bestimmt.

Stimmt das so einfach?

„codon usage bias“

Amino Acid	Codon	<i>Escherichia coli</i>		<i>Saccharomyces cerevisiae</i>	
		High	Low	High	Low
Leucine	UUA	1%	20%	8%	25%
	UUG	1%	15%	89%	25%
	CUU	2%	12%	0%	12%
	CUC	3%	11%	0%	9%
	CUA	1%	5%	3%	15%
	CUG	92%	37%	0%	14%
Valine	GUU	60%	27%	52%	28%
	GUC	2%	25%	48%	19%
	GUA	28%	16%	0%	30%
	GUG	10%	32%	0%	23%
Isoleucine	AUU	16%	46%	42%	43%
	AUC	84%	37%	58%	22%
	AUA	0%	17%	0%	35%
Phenylalanine	UUU	17%	67%	10%	69%
	UUC	83%	33%	90%	31%

**Synonyme
Kodons werden
unterschiedlich
häufig
verwendet**

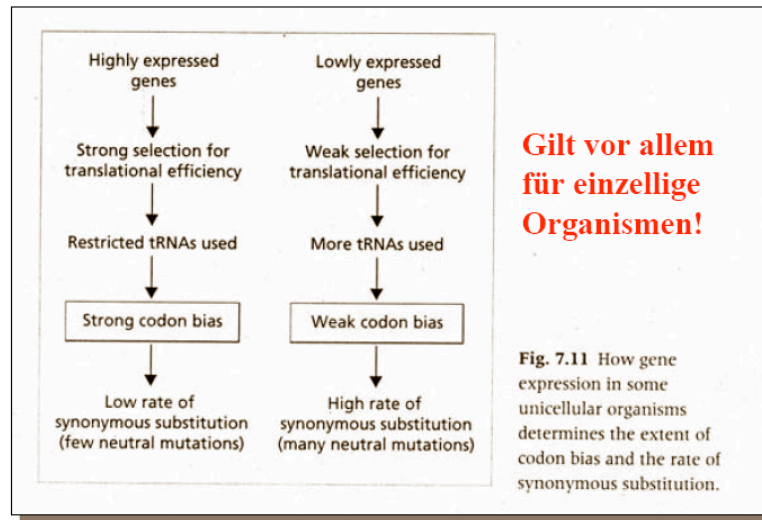


„codon usage bias“

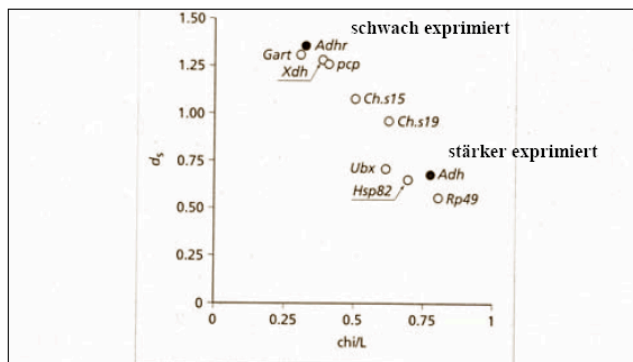
Der Gebrauch synonymer Kodons bestimmt in einzelligen Organismen die Effizienz der Translation und unterliegt daher der reinigenden Selektion!

Hier sind synonyme Mutationen also nicht unbedingt neutral !!

Starke Kodonpräferenz bedeutet langsame Evolution



Starke Kodonpräferenz bedeutet langsame Evolution



... aber auch bei Drosophila!

Fig. 7.12 Relationship between the number of silent substitutions per site (d_s) and synonymous codon bias (χ/L) in 10 *Drosophila* genes. Each point represents the comparison of one gene from the *D. melanogaster* species group (usually *D. melanogaster*) with one from the *D. obscura* species group (usually *D. pseudoobscura*). χ/L represents a χ^2 for deviation from equal use of synonymous codons, divided by gene length (L), averaged for each species pair. *Adh* and *Adhr* are highlighted by closed circles. Adapted from Sharp and Li (1989).

Kodonpräferenz in Säugern

Table 10. Codon usage in four species.*

Amino acid	Codon	<i>Escherichia coli</i>		<i>Saccharomyces cerevisiae</i>		<i>Drosophila melanogaster</i>		Human	
		High	Low	High	Low	High	Low	G + C	A + T
Leu	UUA	0.06	1.24	0.49	1.49	0.03	0.62	0.05	0.99
	UUG	0.07	0.87	5.34	1.48	0.69	1.05	0.31	1.01
	CUU	0.13	0.72	0.02	0.73	0.25	0.80	0.20	1.26
	CUC	0.17	0.65	0.00	0.51	0.72	0.90	1.42	0.80
	CUA	0.04	0.31	0.15	0.95	0.06	0.60	0.15	0.57
	CUG	5.54	2.20	0.02	0.84	4.25	2.04	3.88	1.38
Val	GUU	2.41	1.09	2.07	1.13	0.56	0.74	0.69	1.32
	GUC	0.08	0.99	1.91	0.76	1.59	0.93	1.03	0.69
	GUA	1.12	0.63	0.00	1.18	0.06	0.53	0.11	0.80
	GUG	0.40	1.29	0.02	0.93	1.79	1.80	2.78	1.19
Ile	AUU	0.48	1.38	1.26	1.29	0.74	1.27	0.45	1.60
	AUC	2.51	1.12	1.74	0.66	2.26	0.95	2.43	0.76
	AUA	0.01	0.50	0.00	1.05	0.00	0.78	0.12	0.64
Phe	UUU	0.34	1.33	0.19	1.38	0.12	0.86	0.27	1.20
	UUC	1.66	0.67	1.81	0.62	1.88	1.14	1.73	0.80
Met	AUG	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

From Sharp et al. (1988).

* For each group of synonymous codons, the sum of the relative frequencies equals the number of codons in the group. For example, there are six codons for leucine, and so the sum of the relative frequencies for these six codons should be 6. Under equal usage, the relative frequencies for each codon in a group should be 1, and so the degree of deviation from one indicates the degree of bias in usage. "High" and "low" denote genes with high and low levels of expression. For humans, "G + C" means high-GC regions, and "A + T" means high-AT regions.

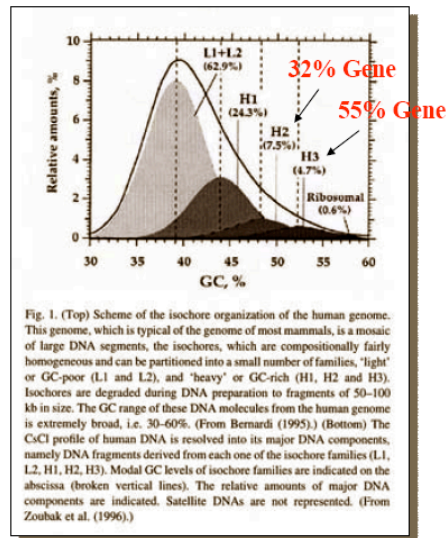
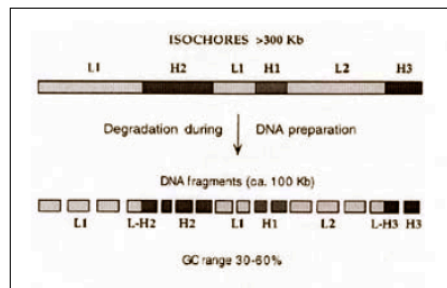
Kodonpräferenz in Säugern

Nicht der tRNA-Pool und die Translationseffizienz bestimmen in Säugern die Kodon-Präferenz,

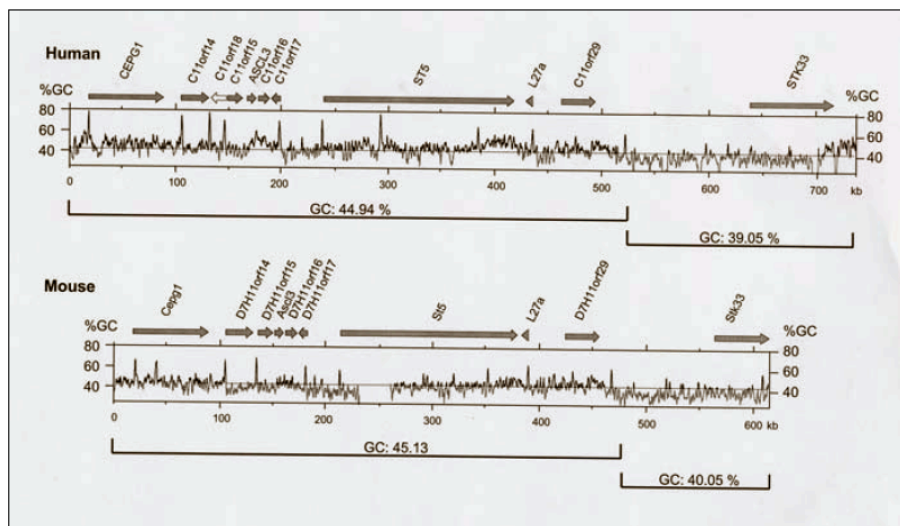
sondern **der AT/GC-Gehalt der Umgebung** („Genomkontext“):

Frage: Ist der AT/GC-Gehalt durch Mutation oder durch Selektion entstanden?

Das Genom warmblütiger Vertebraten hat „Isochoren“ (Bernardi 1985)



Isochoren



Isochoren und Chromosomen-Morphologie

Table 7.2 The properties of isochores in vertebrates. Adapted from Holmquist and Filipski (1994), with permission.

G + C-rich isochores	A + T-rich isochores
Correlate with reverse Giemsa bands (R bands) on chromosomes	Correlate with Giemsa bands (G bands) on chromosomes
Early replicating	Late replicating
High density of genes, including both housekeeping and tissue-specific	Low frequency of genes, only tissue-specific
SINEs present	LINEs present
CpG islands in genes	No CpG islands
High G + C content at third codon position	High A + T content at third codon position
High frequency of retroviral sequences	Low frequency of retroviral sequences
High frequency of chiasmata	Low frequency of chiasmata

More details about the evolution of genome organisation are given in Chapter 3.

Isochoren und codon usage

Isochoren bestimmen AT/GC-Gehalt an synonymen Kodonpositionen und damit die „codon usage“

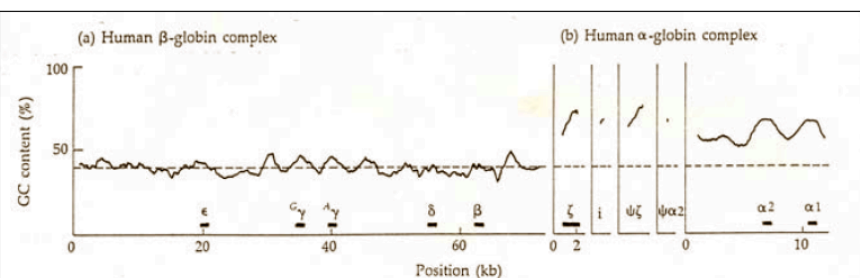


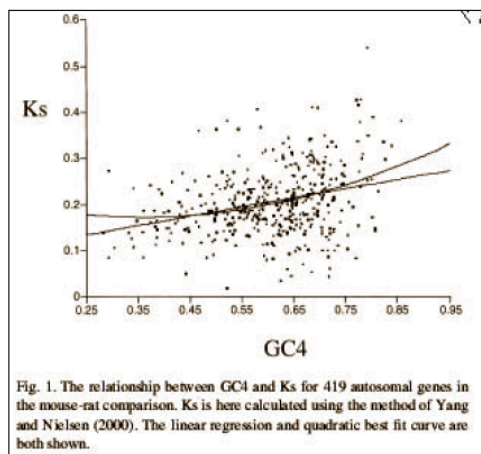
Figure 8. Distribution of GC content along human globin DNA sequences: (a) The β -globin gene cluster; (b) the α -globin gene cluster (incomplete). The genes (bars) are arranged in the same order as in Figure 7 of Chapter 6. The gene names are shown at the bottom of the figure; region i is the intergenic region between ζ and $\psi\zeta$. In the β -globin cluster and the region covering the $\alpha 1$ - and $\alpha 2$ -globin genes each point represents the average of the GC composition of the 2,001 nucleotides surrounding the point, while in the other regions each point represents the average of 1,401 nucleotides. The horizontal broken line represents the overall GC content of the human genome (40%). Modified from Ikemura and Aota (1988).



Sind Isochoren nur durch Mutationspräferenzen entstanden, oder unterliegen sie der Selektion???

Eyre-Walker and Hurst, Nature Rev. Genet. 2001, 2: pp.549

Ein GC-reicher Genomkontext korreliert mit einer erhöhten synonymen Substitutionsrate



Sind die Mutationen von C^{mp}G > TpG in GC-reichen Bereichen verantwortlich?

ca. Faktor 2 Unterschied in dS-rate

Gründe für die Variabilität der synonymen Substitutionsrate

1. Fitness-Unterschiede synonymer Kodons (Selektion)
2. Genomkontext bestimmt Mutationsrate und -richtung