12. Stabilität und Zunahme biologischer Information

Beobachtungen aus Reagenzglas-Evolution (Bakterien, Viren) und aus ALife-Experimenten lassen sich z.T. theoretisch untermauern

typisches Szenario in Experimenten mit genomischer Evolution:

- Genom mit (aktuell) höchster Fitness dominiert die Population
- durch Mutationseinfluss bildet sich Cluster nahe verwandter Sequenzen "um dieses Genom herum" (im Sinne genetischer Distanz) – "*Quasispezies*" (Begriff von M. Eigen)
- wenn Mutationsrate zu hoch: Quasispezies zerfällt, Information kann nicht dauerhaft gespeichert werden: "Fehlerkatastrophe"

2 widerstreitende Anforderungen:

- um mehr Information aus der Umwelt zu speichern (z.B. Lösung komplexer Aufgaben), müssen die Genom-Strings *länger* werden
- um die Information erfolgreich in die nächste Generation weitergeben zu können, dürfen sie *nicht zu lang* werden (Informationsweitergabe muss genau genug sein)

Sei ε_{best} die Selbstreplikationsrate des fittesten Genoms, $\langle\varepsilon\rangle$ die durchschnittliche Selbstreplikationsrate in der Population.

"Superiorität" der besten Sequenz: $s_{best} = \frac{\varepsilon_{best}}{\langle\varepsilon\rangle}$.

Bei gegebener Mutationsrate R (pro Position im Genom) und Genomlänge L ist die Wahrscheinlichkeit einer erfolgreichen Kopie der besten Sequenz: $F = (1 - R)^L$.

Der Erwartungswert der Nachkommenzahl ist dann:

$$E = s_{best} \cdot F = s_{best} \cdot (1-R)^L$$

(Herleitung siehe Adami 1999, S. 278 f.)

Die Sequenz stirbt aus, wenn $E \leq 1$ ist, oder, gleichwertig, wenn gilt:

$$L \geq \frac{-\ln s_{best}}{\ln(1-R)} \approx \frac{\ln s_{best}}{R} =: L_T$$

L_T heißt die *Eigen'sche Fehlerschranke*.

Beispiele aus der realen Biologie:

einfache RNA-Replikation "in vitro": $R \approx 5$ %, $L_T = 10$ bis 100
katalysierte Phagen-Replikation: $R \approx 0{,}01$ %, $L_T = 1000$ bis 10000
Replikation in Bakterien (mit Reparaturenzymen): $R \approx 0{,}00001$ %, $L_T = 10^6$ bis 10^7
Eukaryonten: noch besser

Auflösung nach der kritischen Mutationsrate, wenn Genomlänge L vorgegeben:

$$R < \frac{\ln s_{best}}{L}$$

Wenn Insertions- und Deletionsmutationen berücksichtigt werden (Wahrscheinlichkeit P), ist die Fehlerschranke zu modifizieren (Adami 1999, S. 293):

$$L_T = \frac{\ln s_{best}}{R} - \frac{2P}{R(1-R)^L}$$

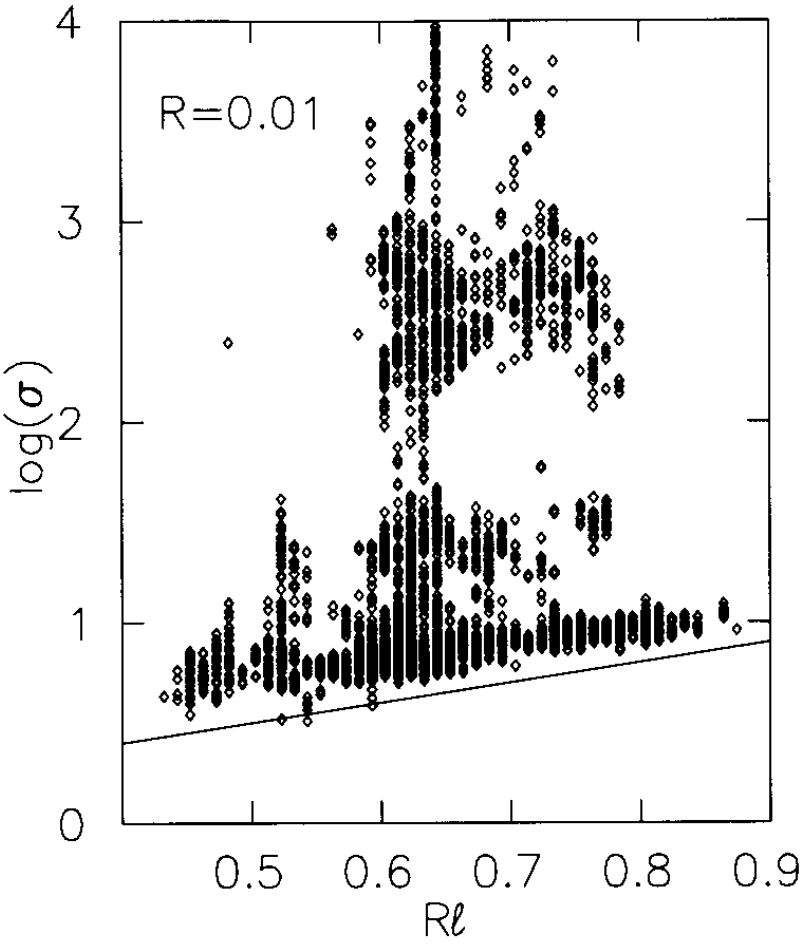
2 Regimes:

- Mutationsrate unter der krit. Grenze: Quasispezies bleibt (vorläufig) erhalten
- Mutationsrate darüber: Quasispezies geht verloren; was danach passiert, hängt von der Fitnesslandschaft ab

Bestätigung der Fehlergrenze in Avida-Testläufen:

- zahlreiche Läufe mit fester Mutationsrate
- 50 000 Update-Schritte; Annahme: danach Gleichgewichtszustand erreicht
- Durchschnittswerte von L und log s_{best} für die letzten 1000 Schritte werden ermittelt (scheiden aus, wenn Varianzen zu groß – dann kein Gleichgewicht)

Ergebnis:



(aus Adami 1999)

(durchgezogene Linie = theoretische Fehlerschranke)

Anpassungsfähigkeit einer Population:

"adaptability" = Rate, mit der die Population Information aus ihrer Umgebung aufnimmt und in biologische (Genom-) Information umsetzt
z.B. bei schrittweiser Lösung einer bestimmten Aufgabe.

auch "Lernrate" genannt (Vorsicht: evolutives "Lernen" gemeint, kein Lernen des Individuums!)

Unter welchen Bedingungen ist die Anpassungsrate maximal?

Maß für die Anpassungsrate:
"learning fraction" = über größere Anzahl von Testläufen gemittelter Anteil "erfolgreicher" Läufe, wenn diese nach fester "cutoff time" abgebrochen werden.

$$f_X(R) = \frac{m}{n}$$

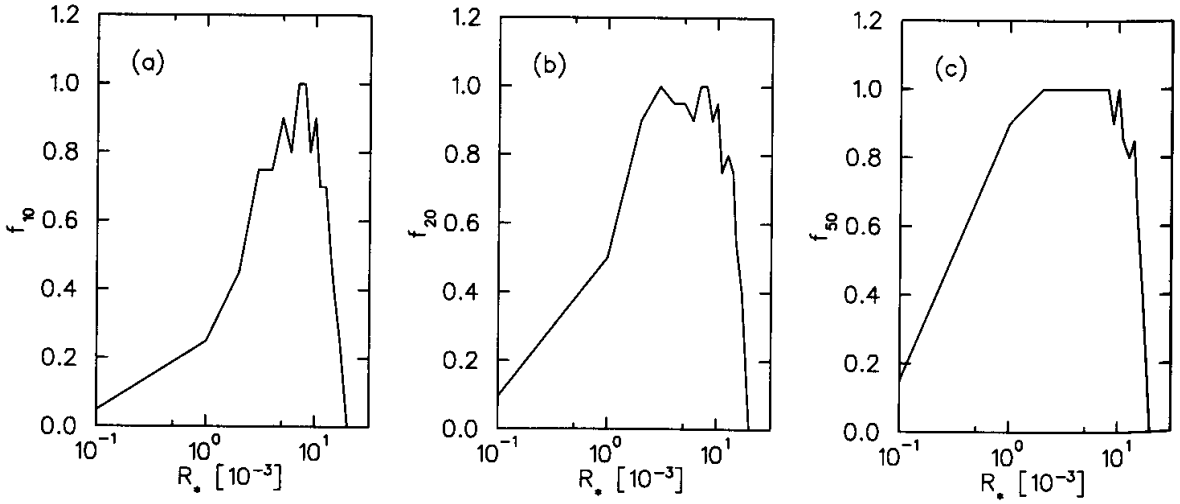
(X = cutoff time, R = Mutationsrate, n = Anzahl der Testläufe, m = Anzahl erfolgreicher Läufe)
Eine Aufgabe gilt als erfolgreich gelöst zu dem Zeitpunkt, wenn der dominierende Genotyp diese Aufgabe zum ersten Mal beherrscht.

Beachte: Wenn es allein copy-Mutationen gibt, kann die Population auch noch bei Raten weit über dem Optimum eine Aufgabe (gelegentlich) lösen (Mutter-Programm funktioniert weiter, auch wenn alle Töchter durch Mutation "verdorben" sind)
– anders bei Punktmutationen:

Lösung einer Additionsaufgabe unter (ausschließlich) Punktmutationen (cosmic ray mutations) in Avida

verschiedene cutoff-Zeiten: (a) X = 10 000, (b) X = 20 000, (c) X = 50 000 Updates

Mutationsrate *R* gegen Anpassungsrate *f*:



steile rechte Flanke = Fehlerschranke!

Beachte: optimale Anpassungsrate liegt nah bei der Fehlerschranke.

> Adami: "learning happens most effectively if the population is almost but not quite melting from the mutational pressure, a population balanced at the edge of chaos – swayed, but holding on."

Vermutung (Eigen 1971):
die Genome beeinflussen ihre Fitnesslandschaft, in der sie evolvieren, so, dass ihre Anpassungsrate maximal wird.

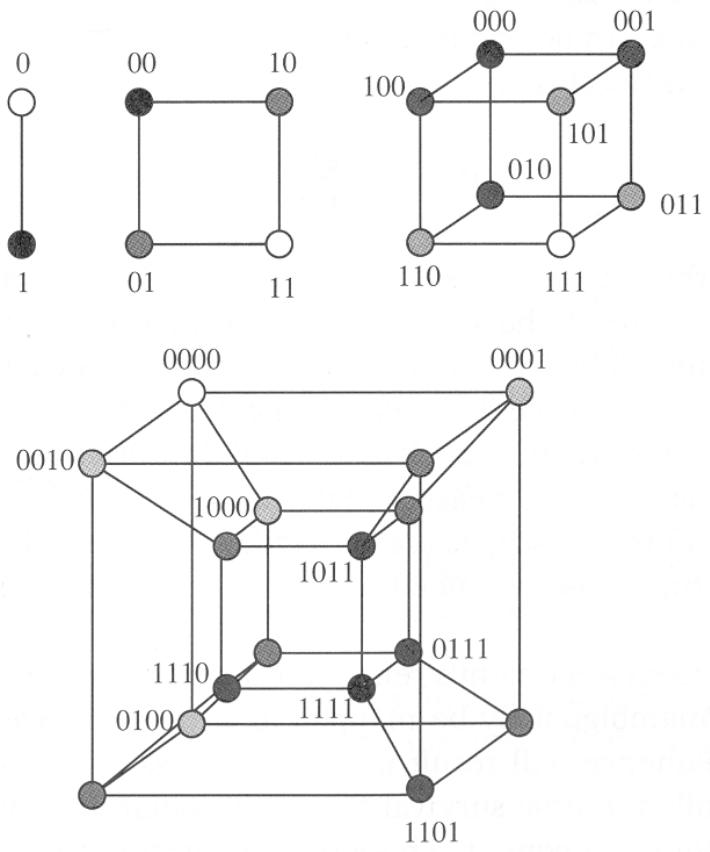
"Evolution to the edge of chaos"

Bestätigung einer Konvergenz zur Fehlerschranke für spezielle, für die Theorie gut zugängliche, feste Fitnesslandschaften (s. Adami S. 285 ff.).

Dynamik der Anpassung hängt stark von der Struktur der Fitnesslandschaft ab.

Wir erinnern:

"Fitnesslandschaft" = Funktion *f* auf Σ^L (Σ = Zeichenvorrat der Genome), im binären Fall auf Hyperwürfel, bzw. auf Σ^*.



Parameter zur Charakterisierung von *f*:

globaler Durchschnitt <f>; Varianz: nicht geeignet, da nicht effektiv bestimmbar (Sequenzraum ist i.allg. zu groß)

Autokorrelation:

$$R(s) = \frac{E(f(x_{t+s})f(x_t)) - E(f(x_{t+s}))E(f(x_t))}{\sqrt{Var(f(x_{t+s}))Var(f(x_t))}}$$

misst die Korrelation zwischen den Fitnesswerten zweier Sequenzen x_t und x_{t+s} , die durch *s* Mutationsschritte getrennt sind
– durch lokale Messungen (Stichproben) schätzbar

Pfad-Autokorrelation:

$$\overline{R}(s) = \frac{\langle f(x_0)f(x_s)\rangle - \langle f(x_0)\rangle^2}{Var(f(x_0))}$$

wobei über alle Pfade, die x_0 und x_s verbinden, gemittelt wird.

Eine Landschaft ist *self-averaging*, wenn $R(s) = \overline{R}(s)$.

Dritte Autokorrelations-Definition:
Abhängigkeit von der Hamming-Distanz *d* (shortest mutational walk):

$$\rho(d) = \frac{\langle f(x)f(y)\rangle_\Delta - \langle f(x)\rangle^2}{Var(f(x))}$$

(Mittelwertbildung über alle Sequenzen *x*, *y* mit Hamming-Distanz *d* voneinander)

Für sehr hochdimensionale Räume führen fast alle Pfade der Länge *s* auch zu Sequenzen der Hamming-Distanz *s* ("few walks backtrack genetically") $\Rightarrow$ dann ist $R(s) \approx \rho(s)$.

Korrelationslänge:
Längenschranke *x*, so dass alle Sequenzen, die größere Hamming-Distanz als *x* haben, (annähernd) unkorrelierte Fitness aufweisen.
(nicht für alle Fitnesslandschaften def.)

Beispiele spezieller Fitnesslandschaften:

- *f* = const.

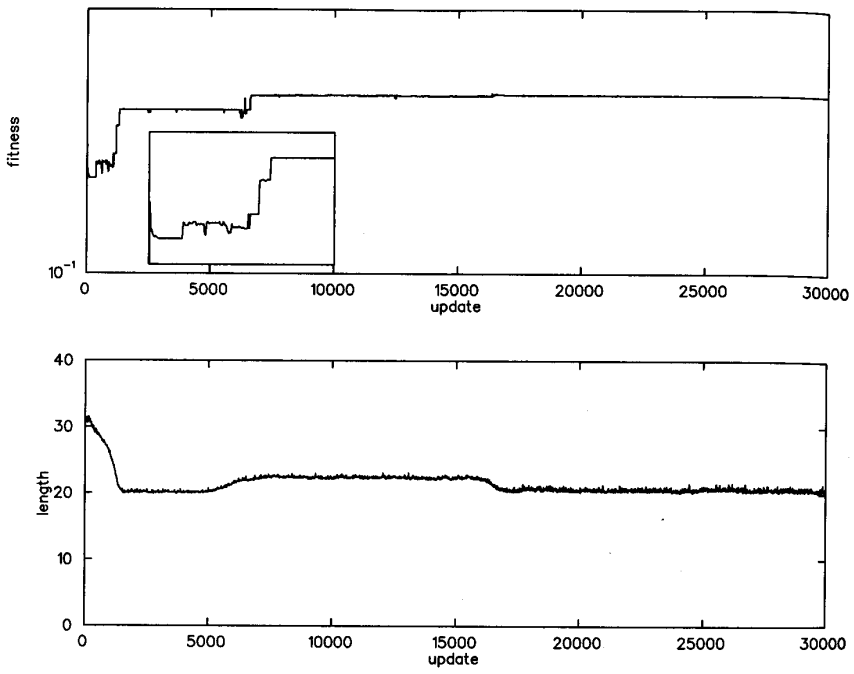
total flache Landschaft, nur neutrale Evolution möglich
(Referenzfall für ALife-Hypothesenprüfung, vgl. LindEvol)

- *f* mit einem globalen Maximum und "glatten Flanken", z.B. *f*(*x*) = Anzahl der Einsen in *x*:

"Fujijama-Landschaft" (Kauffman); Gipfel von überall zu finden

wenn diese Landschaft für *jede* Genomlänge *L* vorliegt, sind kürzere Genome im Vorteil, da sie das Maximum (im kleineren Sequenzraum) schneller erreichen
⇒ *keine* Annäherung an die Fehlerschranke
Landschaft "zu einfach" für interessante Evolution

Beispiel für Evolution in einer (relativ) einfachen Landschaft (nicht ganz ohne Struktur), Avida-Beispiel (oben Fitness, unten Genom-Länge; aus Adami 1999):



Anderes Extrem:

- $f(x)$ Zufallsvariable, unabhängig von jedem Nachbarstring y. $\rho(0) = 1$; $\rho(1) = \rho(2) = ... = 0$.

"*Derrida-Landschaft*", random energy model.
Total zerklüftete Landschaft.

Hier ist keine Evolution (durch schrittweise Verbesserung) möglich!

Objekte, die diesem Fall nahekommen:

- kürzeste Rechnerprogramme, kürzeste Formeln

Beschreibungen ohne jegliche Redundanz
$\Rightarrow$ kleine Änderungen im Code führen zu Objekten mit völlig anderer Semantik
$\Rightarrow$ schwere Evolvierbarkeit dieser Objekte mit GA (vgl. Beispiel der Binet'schen Formel für Fibonacci-Zahlen!)

$\Rightarrow$

Notwendigkeit von Redundanzen für die Evolution

vgl. Redundanzen in den Befehlssätzen von Avida, Tierra; Introns in der DNA...

Modell-Fitnesslandschaften, mit denen der Zerklüftungsgrad zwischen den beiden Extremen (Fujijama- und Derrida-Landschaft) getunt werden kann:

NK-Fitnesslandschaften (S. Kauffman)

siehe Kauffman 1995, S. 260 ff.

- die *N* einzelnen Symbole im Genom-String werden vereinfachend mit "Genen" gleichgesetzt
- jedes Gen liefert einen "Fitnessbeitrag" für das gesamte Genom
- Fitness des Genoms der Länge *N* = Mittelwert der *N* Beiträge seiner Gene
- Annahme: der Fitnessbeitrag jedes Gens wird noch von der Ausprägung von *K* anderen Genen beeinflusst (epistatische Kopplung; vgl. Kap. 10)

Festlegung einer *NK*-Fitnesslandschaft (wenn *N* und *K*, $0 < K < N$, vorgegebene Zahlen sind):

- lege zu jedem der *N* Gene zufällig *K* andere Gene als "Inputgene" fest
- für jedes Gen *x*:
  lege (stochastisch unabhängig) für jede Allelkombination, die an den *K*+1 beteiligten Genen (*x* und seine Inputgene) auftreten kann, eine zufällige Zahl (gleichverteilt aus [0; 1]) als Fitnessbeitrag von *x* fest

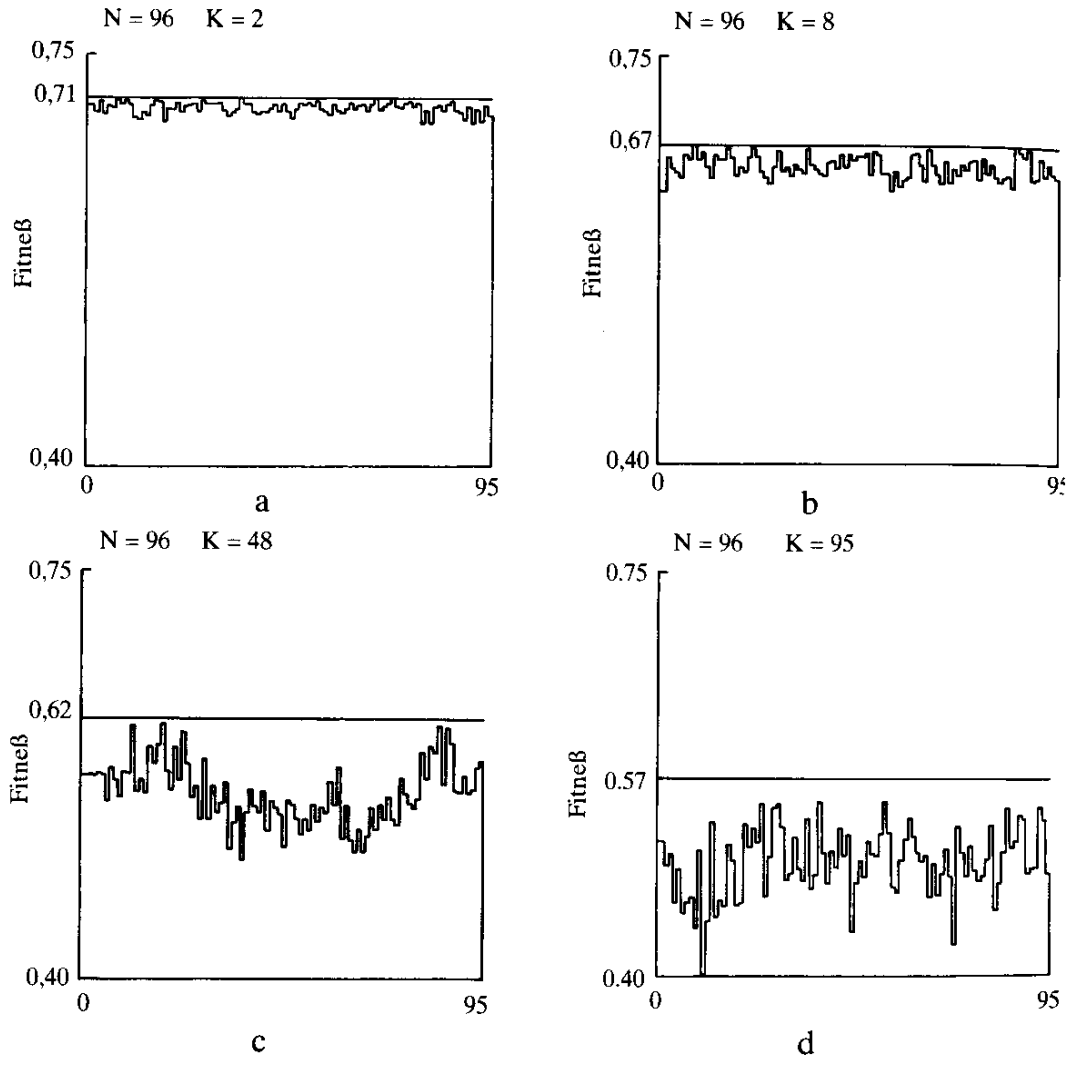
beachte: bei vielen Kopplungen (großes *K*) kommt es zu vielen "widerstreitenden Bedingungen" (eine Kombination von Genen kann als Input für Gen *x* dessen Fitness senken, als Input für *y* dessen Fitness erhöhen...)

Extremfälle:

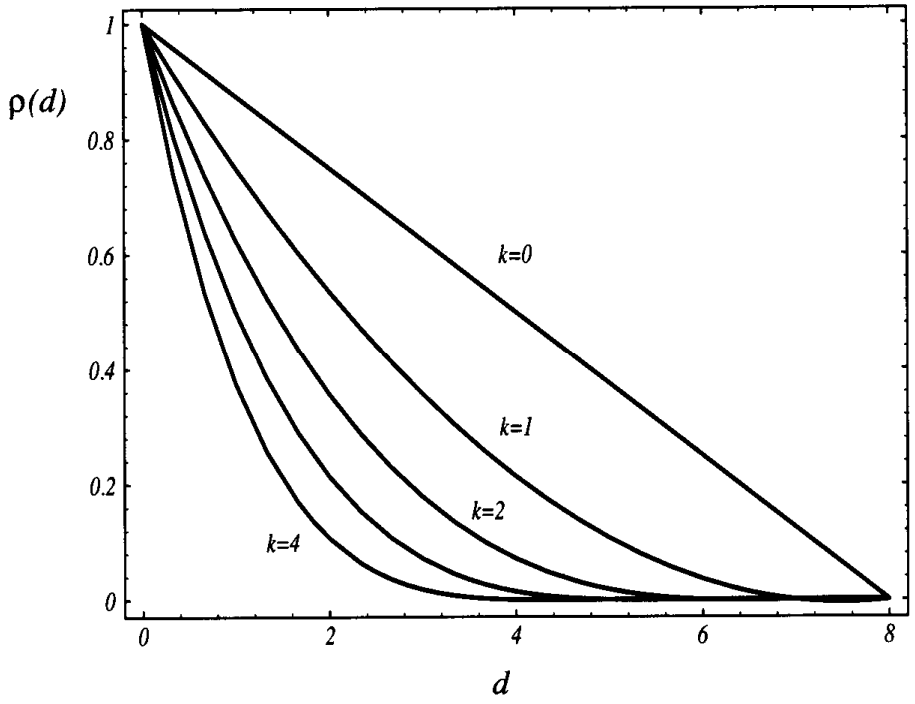
- $K = 0$: keine epistatische Kopplung, jedes Gen liefert seinen unabhängigen (additiven) Beitrag zur Fitness $\Rightarrow$ "Fujijama-Landschaft"
- $K = N - 1$: alle Gene sind mit allen gekoppelt, Fitness des ganzen Genoms ist für jede Allelkombination stochastisch unabhängig festgelegt $\Rightarrow$ Derrida-Landschaft

dazwischen: je größer K, desto zerklüfteter die Landschaft

Darstellung der Nachbarschaft eines lokalen Optimums für 4 verschiedene Werte von K (aus Kauffman 1995):



Autokorrelationsfunktion im *NK*-Modell ($N = 8$) (aus Adami 1999):



- bei niedrigen *K*-Werten: Landschaft nichtisotrop, Gipfel häufen sich in spezieller Region
  $\Rightarrow$ für adaptiven Prozess ist es vorteilhaft, diese Region aufzusuchen
- bei hohen *K*-Werten: Isotropie, hohe Gipfel sind zufällig verstreut

Merkmal mittelstark zerklüfteter *NK*-Landschaften:

die höchsten Gipfel können von der größten Zahl an Ausgangspositionen erklommen werden (mit Gradienten-methode, d.h. ohne zwischendurch abzusteigen)

Bei Invertierung der Landschaft (Gipfel $\rightarrow$ Täler) würde man sagen: die tiefsten Täler entwässern die ausgedehntesten Einzugsgebiete

Beispiel $N = 24$, aufgetragen ist Anzahl der Ausgangspositionen für ein lokales Maximum gegen dessen Fitness-Wert (aus Kauffman 1995):

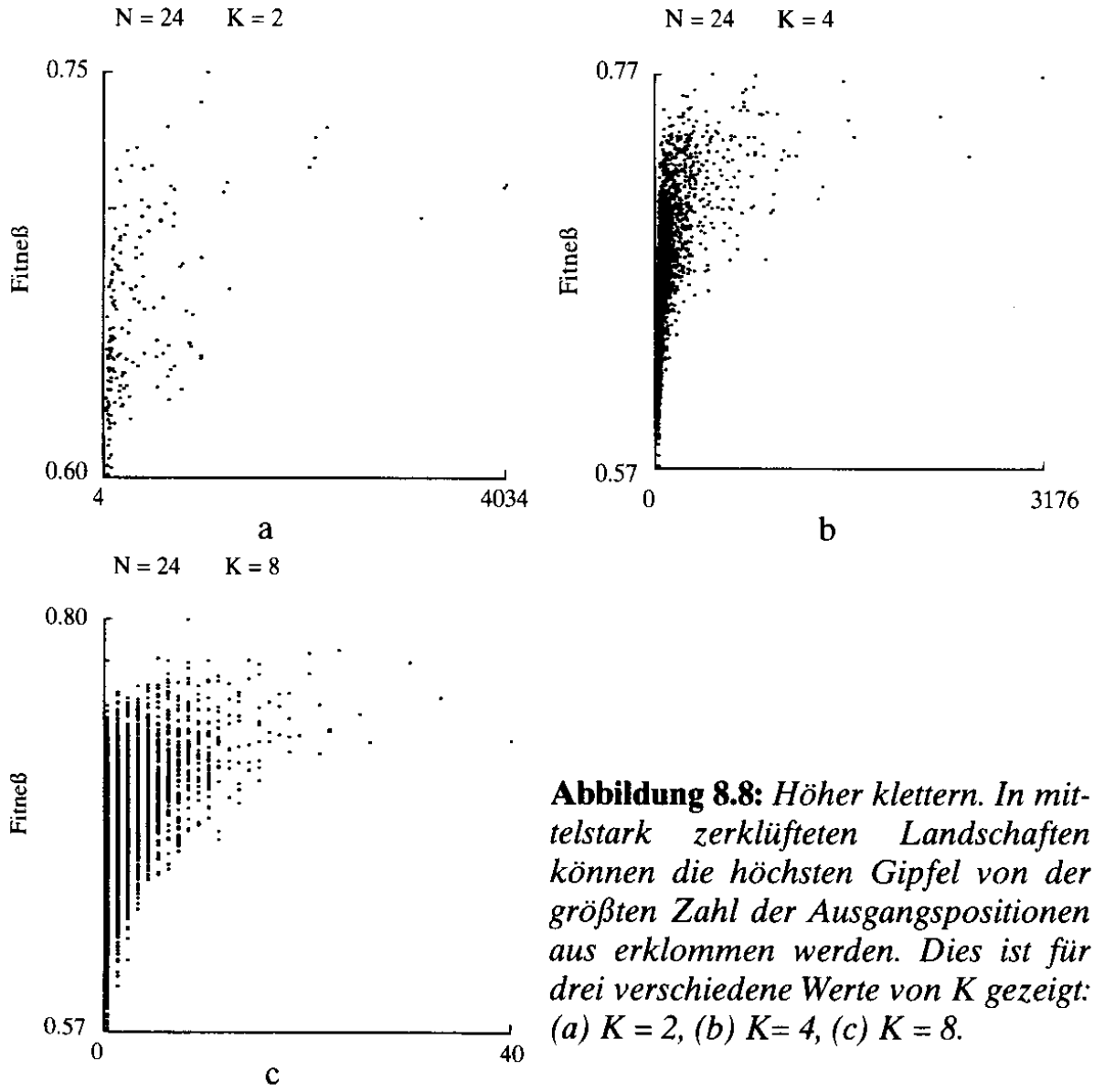


Abbildung 8.8: *Höher klettern. In mittelstark zerklüfteten Landschaften können die höchsten Gipfel von der größten Zahl der Ausgangspositionen aus erklommen werden. Dies ist für drei verschiedene Werte von K gezeigt: (a) K = 2, (b) K= 4, (c) K = 8.*

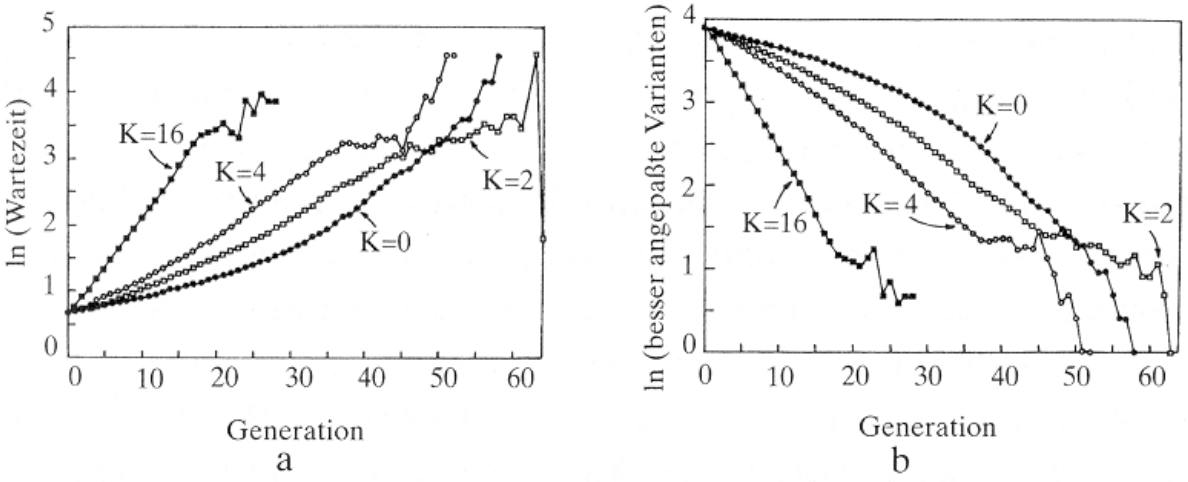
⇒ "günstig" für Evolution!

Wie verhält sich die Anzahl *b* der benachbarten Genome höherer Fitness während eines Aufstiegs?

- *K*=0 (Fujijama-Situation): lineare Abnahme
- für mittelstark bis stark zerklüftete *NK*-Landschaften (ab ca. *K* = 8): Abnahme um konstanten Faktor

⇒ im letzteren Fall nimmt Anzahl der nötigen Versuche, um ansteigenden Pfad zu finden, um konst. Prozentsatz zu

"Weg wird am Gipfel immer beschwerlicher"
"Abnehmende Erträge" der Anpassung

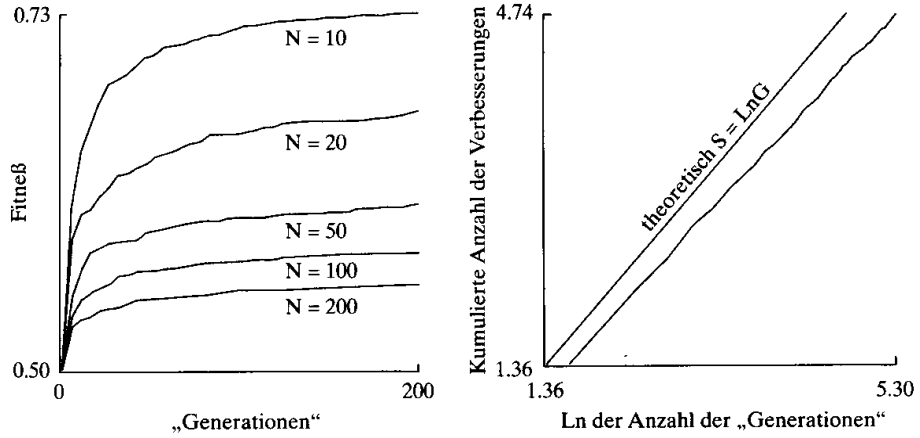


⇒ exponentielle Verlangsamung der evolutiven Optimierung, beobachtbar bei natürlicher und bei technologischer Evolution

- diese Betrachtungen galten für Wanderungen mit "kleinen Schritten" in der Landschaft (adaptive walks)
- es sind auch größere Sprünge denkbar (Mutationen an mehreren Genen gleichzeitig)
- "Weitsprünge": Distanz größer als die Korrelationslänge

hier gilt ebenfalls das exponentielle Verlangsamungsgesetz (schon für kleinere *K*):
jedesmal, wenn man eine "Weitsprung"-Variante höherer Fitness gefunden hat, verdoppelt sich die Anzahl der Versuche, um eine noch bessere Weitsprung-Variante (vom neuen Ort aus) zu finden

Beispiel $K = 2$: (x-Achse: kum. Anzahl unabh. Weitsprungversuche)



aus Bild a wird auch noch ersichtlich: bei größerer Genomlänge *N* erzielen Weitsprung-Adaptationen immer schlechtere Ergebnisse.

- "Komplexitätskatastrophe": je komplexer (länger) das Genom, umso schwieriger ist es, tiefgreifende vorteilhafte Veränderungen anzuhäufen.

Aus den Beobachtungen zur Anpassungsdynamik in *NK*-Landschaften (mit mittlerem *K*) folgt:

"3 Phasen der Evolution"

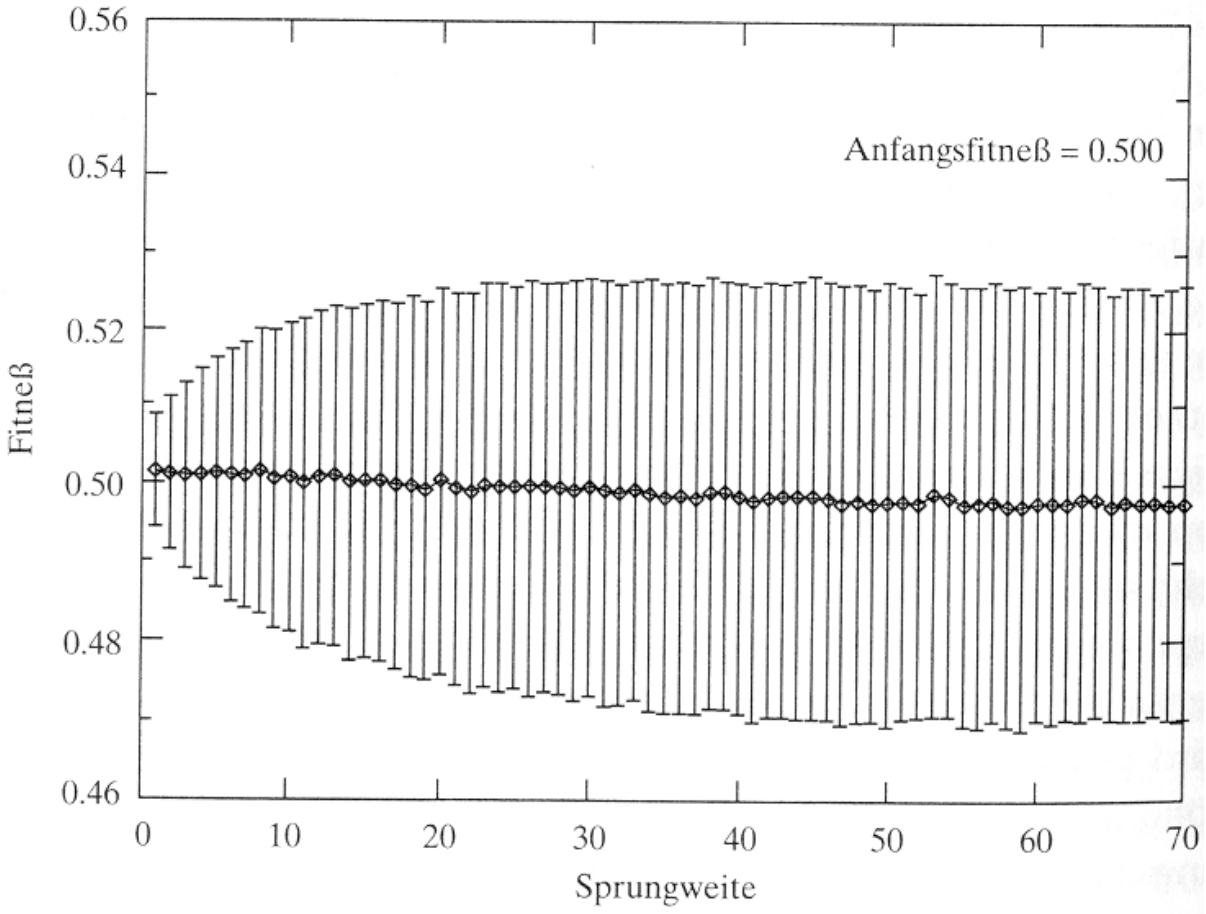
Die Ausgangsposition habe eine mittlere Fitness.

1. Phase: in der engen Nachbarschaft gibt es nur geringfügig höhere Fitnesswerte, in Entfernungen jenseits der Korrelationslänge aber evtl. Stellen mit sehr viel höherer Fitness
$\Rightarrow$ wenn einige Adaptationsvarianten die Umgebung absuchen und einige Weitsprünge vollführen, sind in dieser Phase die Weitsprünge bald erfolgreich (in der Regel sogar mehrere ganz verschiedene Varianten!)
$\Rightarrow$ *Radiation* (Entstehung zahlreicher ganz neuer Baupläne; vgl. kambrische Explosion)

2. Phase: Wartezeit für bessere Weitsprung-Varianten nimmt exponentiell zu; Prozentsatz der besseren Varianten in der Nähe nimmt aber langsamer ab ⇒ Evolution macht nur noch kleine Sprünge (Verfeinerung der bestehenden Baupläne)

3. Phase: Evolution festgefahren auf lokalen Gipfeln, oder Wanderung entlang Graten hoher Fitness, oder Landschaft ändert ihre Form und Populationen folgen den sich verschiebenden Gipfeln

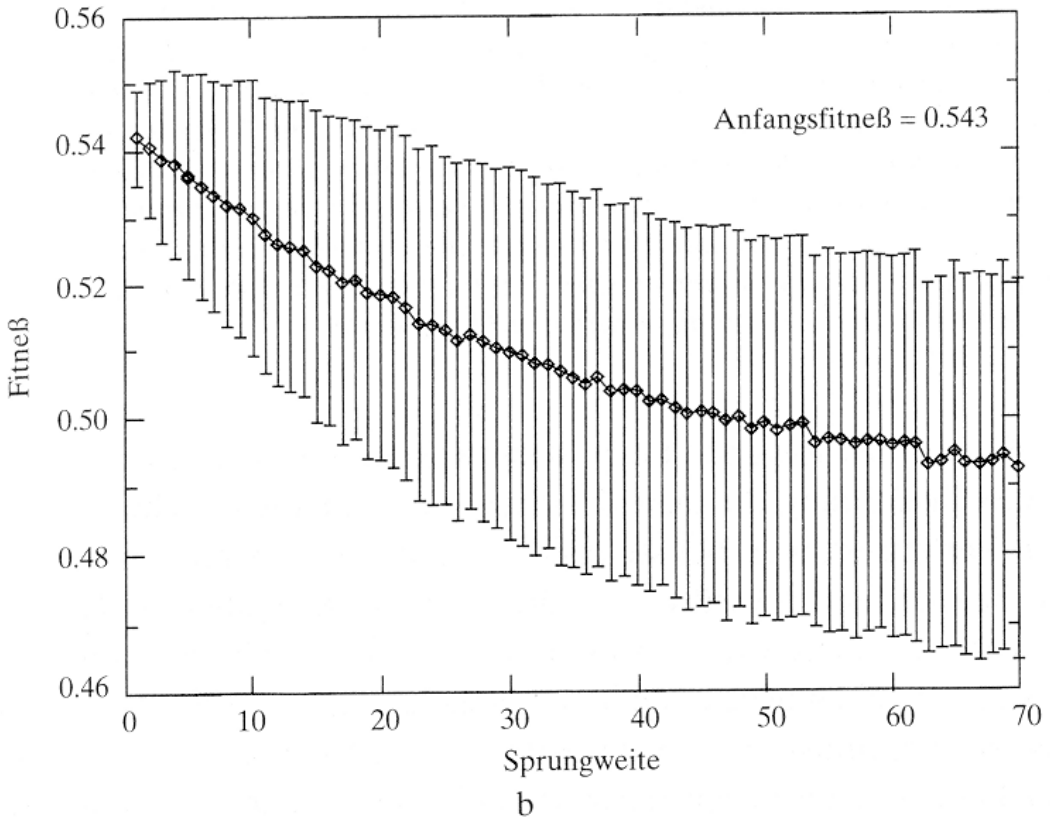
Beziehung zwischen Sprungweite und erreichbarer Fitness (Balken = Standardabweichung) auf Basis von jeweils 1000 Versuchen
(in korrelierter *NK*-Landschaft; aus Kauffman 1995):



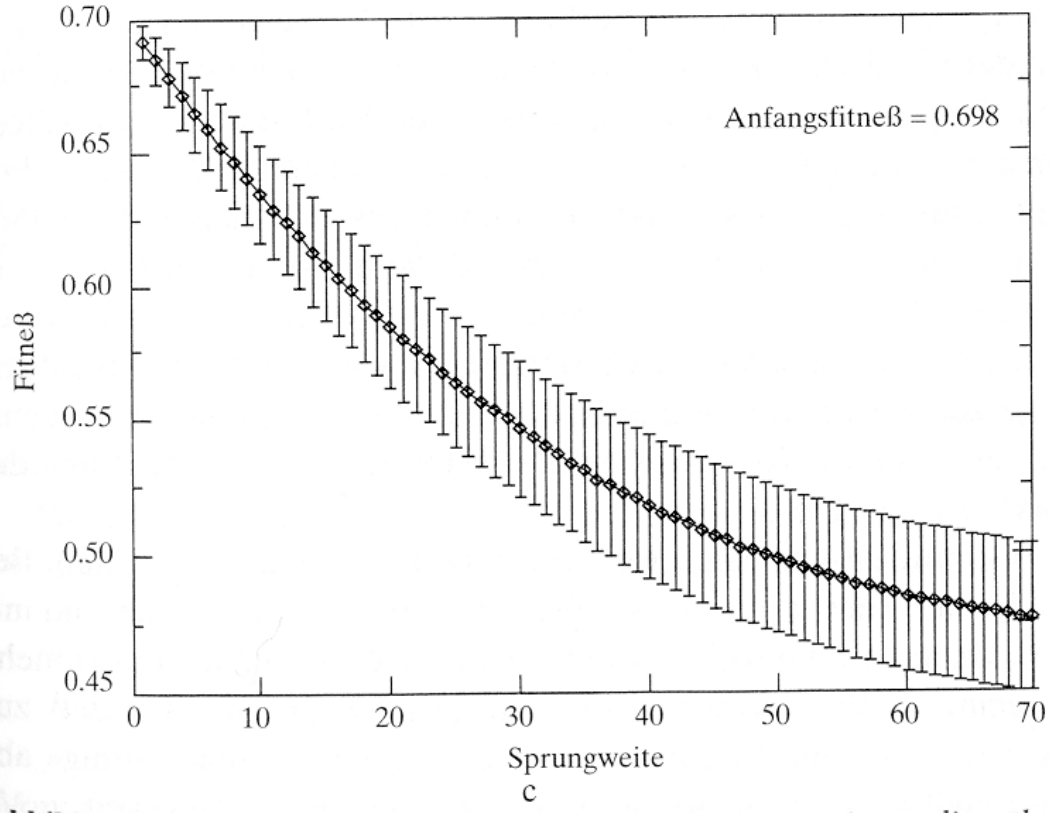
a

bei höherer Ausgangsfitness ändert sich das Bild:

je mehr die Fitness ansteigt, desto ratsamer ist es, die nähere Umgebung abzusuchen



b



c

Evolution durchläuft die Fitnesslandschaft "blind"

"Gottes Perspektive": globale Kenntnis der Fitnesslandschaft – praktisch niemals erreichbar

Bedeutung der Sexualität:

durch *Rekombination* können Genomzustände auf mittleren Distanzen *zwischen 2 Genomen* getestet werden

$\Rightarrow$ ermöglicht Annäherung an eine globalere Sicht

- Rekombination erhöht Adaptationsrate in korrelierten *NK*-Fitnesslandschaften
- nicht in der Derrida-Zufallslandschaft

dass sich Sex weitgehend durchgesetzt hat:

ein Zeichen, dass die Fitnesslandschaften der realen Biologie genügend korreliert sind, so dass Sex von Vorteil ist!

Komplexitätsmaße

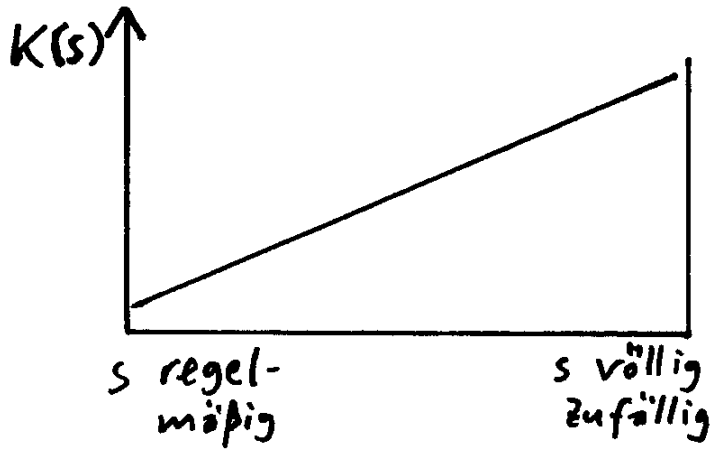
Entropie: statistischer Begriff, nicht anwendbar auf 1 einzelnes Genom

aber: einige Strings erscheinen komplexer (irregulärer) als andere

Kolmogorov-Komplexität
(Kolmogorov 1965; manchmal auch Chaitin-Komplexität, Chaitin 1966):

K(*s*) = Länge des *kürzesten Programms*, das erforderlich ist, um den String *s* zu berechnen (mittels einer universellen Turingmaschine *T*)

regelmäßige Strings *s*, z.B. 0000... oder 010101...: kleines *K*(*s*)
"Zufalls-Strings": großes *K*(*s*); *s* völlig zufällig: *K*(*s*) ≈ Länge(*s*).



K(*s*) als Komplexitätsmaß 1. Ordnung

Nachteile:

- "echten" Zufalls-Strings will man keine große Komplexität zubilligen ("Würfeln" ist nicht sehr kompliziert!)
- Regularität eines Strings sagt noch nichts über Komplexität der Bedeutung aus (Art der Codierung hat Einfluss)

Beisp. (Latein):
TETERROROMAMANUNUDADATETELALATETE
= bedeutungsvoller Satz

notwendig: Einbeziehung eines Kontexts / Bedeutungszusammenhangs / eines Universums

formal: gegebener String *u*, der das "Universum" repräsentiert

Bedingte Komplexität (conditional complexity, Kolmogorov 1983):

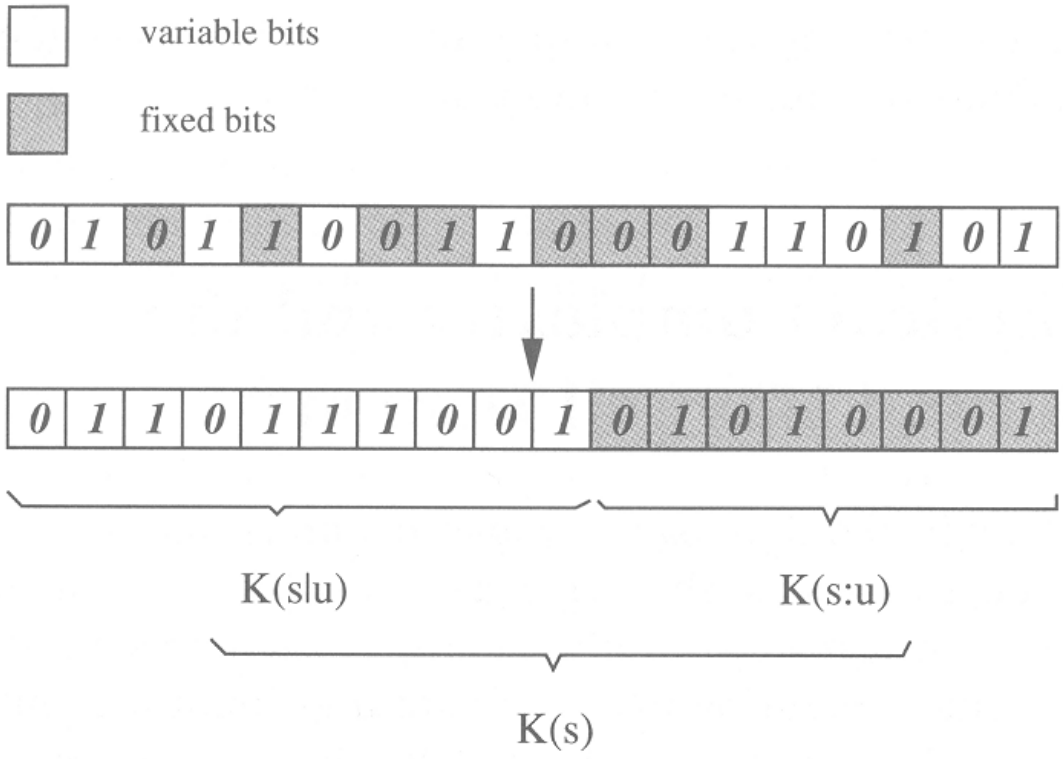
$K(s \mid u)$ = Länge des kürzesten Programms, das *s* berechnet, wenn die Turing-Maschine *u* als Input erhält.

- misst die "verbleibende Regellosigkeit" in *s*, wenn *u* bekannt ist
- zählt diejenigen Bits in *s*, die nicht mit Bits in *u* "korreliert" bzw. durch *u* redundant gemacht sind

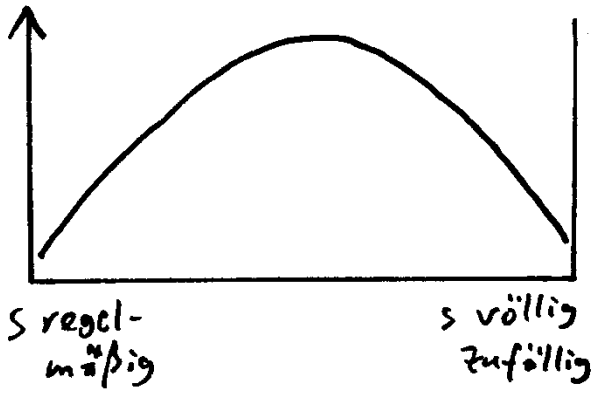
Wechselseitige Komplexität (mutual complexity; auch: physikalische Komplexität, Adami 1998):

$K(s : u) = K(s) - K(s \mid u)$

misst die Anzahl der Bits, "die im Universum *u* etwas bedeuten" bzw. durch Berechnung aus *u* entstanden sein können



K(*s* : *u*) ist ein "Komplexitätsmaß 2. Ordnung":



(wobei "Zufälligkeit" hier nur in Bezug auf *u* definiert werden kann und i.allg. unentscheidbar ist.)

Physikalische Sichtweise:
Gewinnung von *s* aus *u* (mittels einer universellen Turing-maschine) kann als Messvorgang im durch *u* beschriebenen Universum *U* angesehen werden

Zusammenhang zwischen Kolmogorov-Komplexität und Shannon-Entropie:

Durchschnittliche Kolmogorov-Komplexität einer Menge *S* von Strings entspricht der Entropie.

$$H(S \mid u) = \langle K(s \mid u) \rangle_S = -\sum_s p(s \mid u) \log p(s \mid u) ,$$

entspr. für die wechselseit. Information (Shannon)
(Information über *u* in *S*):

$$I(S : u) = H(S) - H(S \mid u) = \langle K(s : u) \rangle_S$$

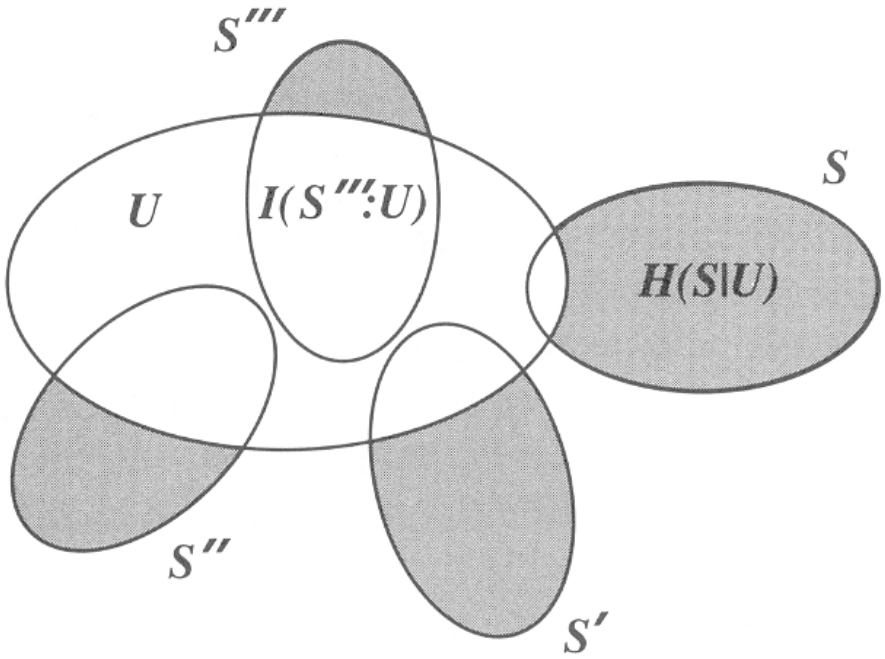
(s. Adami 1998, S. 127 ff.)

Jede Berechnung (Messung) reduziert die bedingte Entropie von *S* und vergrößert die über *U* gewonnene Information, die in der Menge *S* der Strings enthalten ist.

Wenn unter Einfluss von Mutationen ein Bit in *s*, das vorher "zufällig" (in Bezug auf *U*) war, jetzt auf eine Eigenschaft von *U* antwortet ("bessere Anpassungsleistung"), d.h. aus *U* durch Berechnung (Messung) gewonnen werden kann,

so repräsentiert dieses Bit jetzt *Information* (über das Universum *U*).

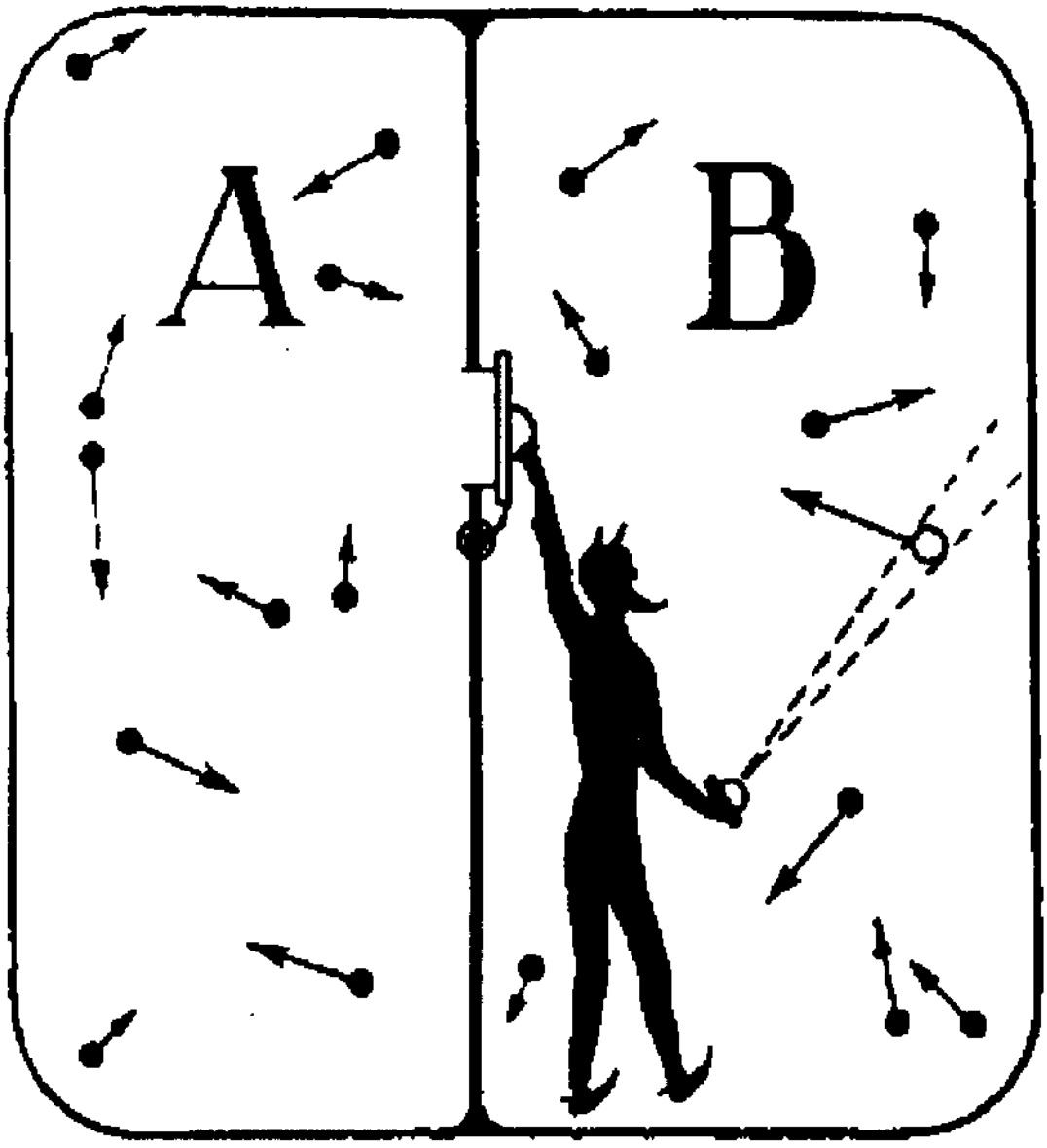
Evolution einer Menge von Strings $S \rightarrow S' \rightarrow S'' \rightarrow S'''$ als Prozess der Abnahme der bedingten Entropie von *S* und der Zunahme der Information über *U*
($H(S) = H(S|U) + I(S:U)$):



- Mutationen als fortlaufende, spontane Messungen
- immer, wenn eine Messung erfolgreich war, nimmt die (konditionale) Entropie der String-Population ab und die Information zu
- "bedeutungsvolle" Bits bleiben in S erhalten aufgrund der Selektion

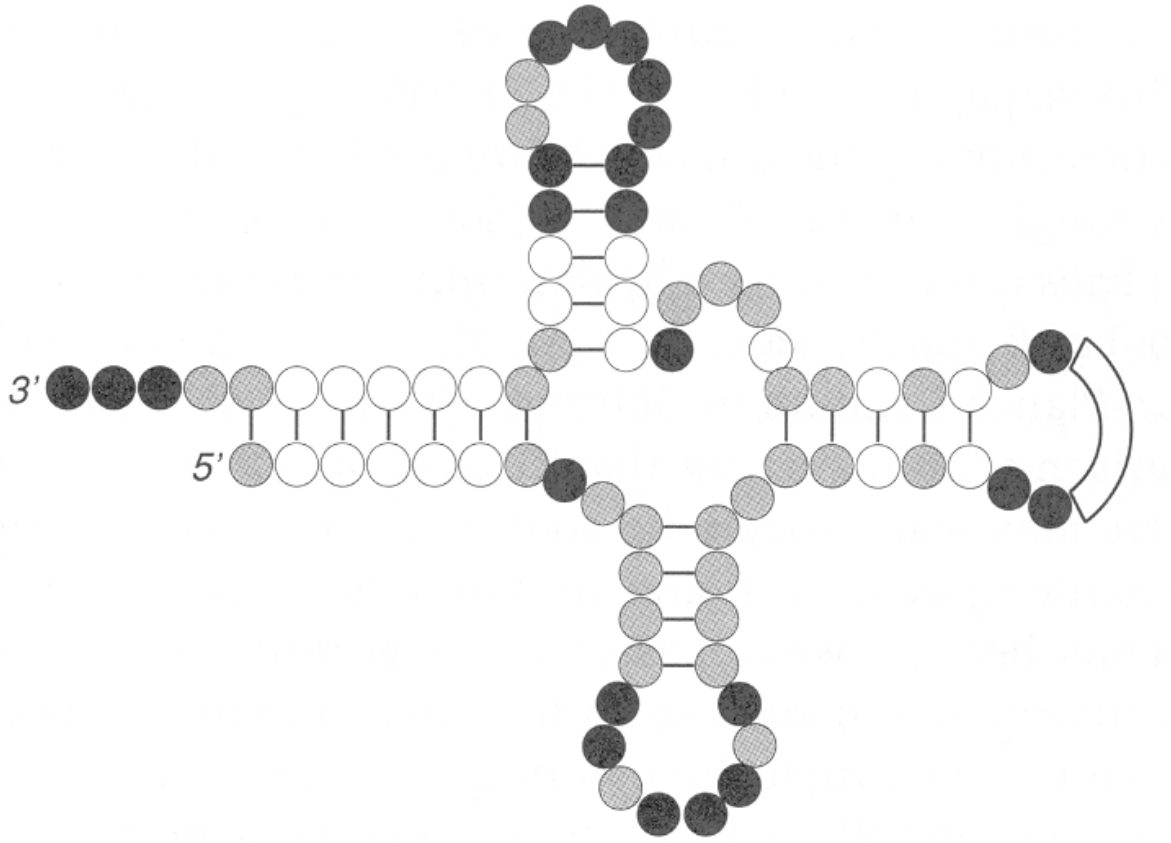
Analogie zum klassischen *Maxwell'schen Dämon* aus der Thermodynamik (Maxwell 1871):

an Elementen einer Population (Teilchen) in einem Behälter werden Messungen durchgeführt; durch Öffen und Schließen einer Klappe gelangen nur "schnelle" (hier: in *U* erfolgreiche) Elemente in die andere Hälfte des Behälters.
Dadurch Abnahme der Entropie im System (Zunahme der Wärmedifferenz zwischen den Hälften *A* und *B*).



- im Unterschied zum klassischen Maxwellschen Dämon (der physikalisch nicht möglich ist) verkleinert der Dämon der Evolution nur die *bedingte* Entropie.

Klassifikation von "informationstragenden" und "zufälligen" Nukleotiden am Beispiel der tRNA (Molekül seit ca. 4 Milliarden Jahren evolviert, in zahlreichen Arten sequenziert):

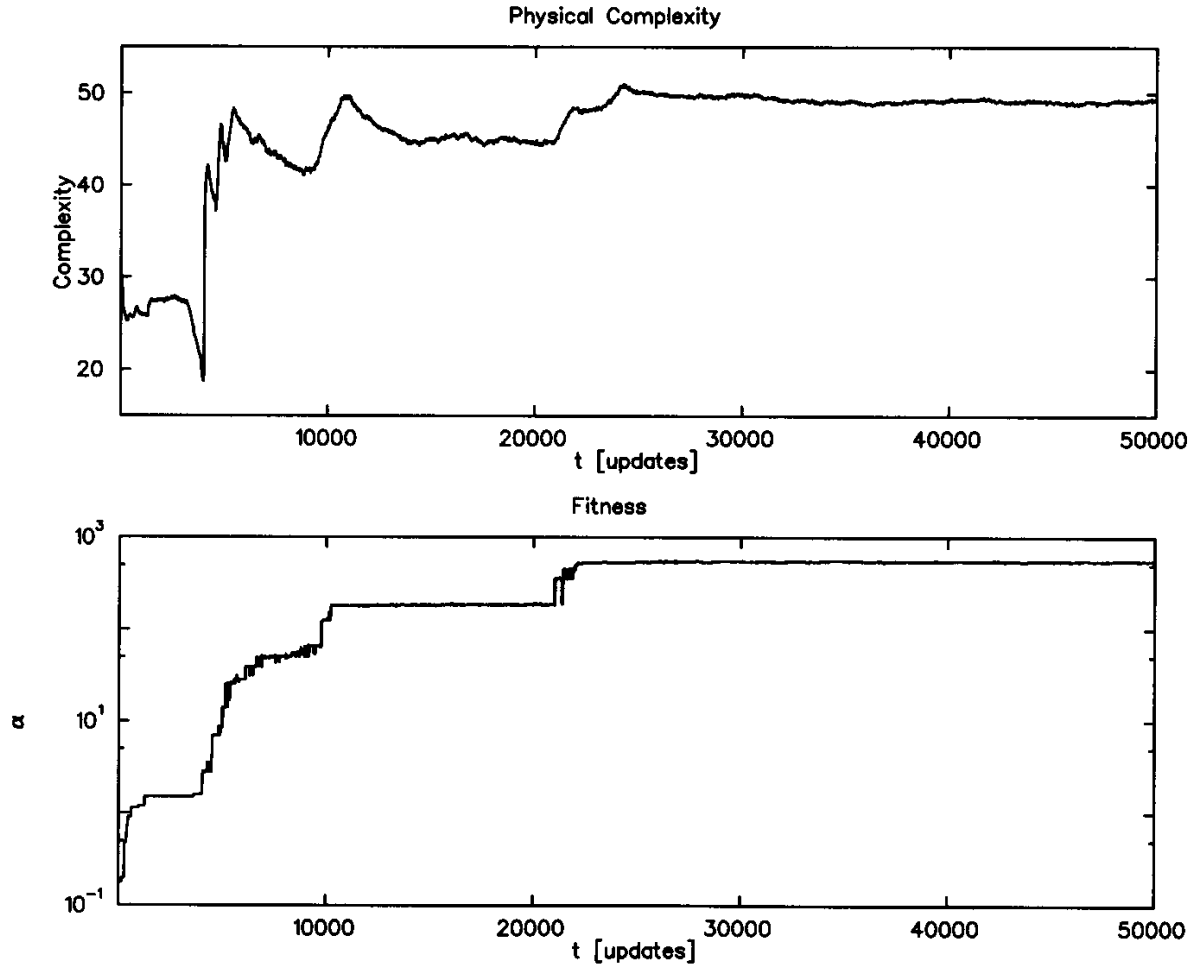


schwarz: unveränderte Positionen
grau: Positionen, die schwacher Veränderung unterliegen
weiß: hoch veränderliche Positionen

(Adami 1998, nach Eigen 1989)

$\Rightarrow$ Komplexität der tRNA liegt zwischen 21 (schwarz) und 42 (schwarz + grau) Bits (wenn nur Basen A/U vs. G/C unterschieden werden)

Entwicklung der physikalischen Komplexität in ALife-Experimenten (hier: Avida; Adami 1998):



(oben: Komplexität, unten: Fitness)