Molekularbiologie und Genetik II (für Studierende der Bioinformatik): Vorlesungsskript

Peter N. ROBINSON
peter.robinson@charite.de

Version vom 17. Mai 2009

Inhaltsverzeichnis

Vorwort

Dieses Skript enthält v.a. Erklärungen zu den Themen Mutation, Selektion, random drift, neutrale Variation, genetische Marker Rekombination, 'lod scores', Kopplungsgleichgewicht und Populationsgenetik in Ergänzung zu den entsprechenden Vorlesungen 'Genetik für Bioinformatiker'.

Kommentare, Vorschläge, Korrekturen usw. bitte an peter.robinson@charite.de

Rechtliche Hinweise

Kapitel 1

Die Organisation des genetischen Materials in den Chromosomen

Ein Verständnis der strukturellen Organisation des genetischen Materials in Chromosomen bzw. der Komplexität eukaryontischer Chromosomestruktur ist wichtig, um die Mechanismen der Evolution, der genetischen Kopplung und der Steuerung der Genexpression zu begreifen. In Ergänzung zum Stoff in Strachan & Read, Kapitel 2, sind hier die wichtigsten Begriffe aufgeführt. Am Schluss des Kapitels finden sich Hinweise über Literatur zu neueren Erkenntnisse über Copyzahlvarianten und verwandten Themen.

Definition 1 (Chromosom) Chromosomen sind Komplexe aus DNA und Proteinen, die unter anderem Gene und regulatorische DNA-Sequenzen enthalten. Der Name "Chromosom" heißt wörtlich Farbkörper und spielt auf die "Anfärbbarkeit" der Chromosomen an (z.B. die Giemsa-Bänderung, welche dunkle G-Banden und helle G-negative Banden in den Chromosomen sichtbar werden lässt. Die Mischung aus DNA und Proteinen in Chromosomen wird auch als **Chromatin** bezeichnet. □

Die wichtigsten Proteine im Chromatin sind die **Histone**, positiv geladene Proteine mit einem Molekulargewicht von etwa 10.000–20.000 Dalton. Die Histone lassen sich in 5 Klassen einteilen, H1, H2A, H2B, H3 und H4. Die Histone sind nicht nur für die Verpackung der DNA wichtig (sie dienen als Teil der Nukleosome gleichsam als Spulen, um die sich die DNA windet), sondern steuern die Genexpression durch ihren Einfluss auf die Chromatinstruktur. Unterschiedliche Modifikationen der Histone (u.a. Methylierung, Phosphorylierung und Azetylierung) beeinflussen die Aktivität der Histone und somit die Expression der Gene. Als Beispiel sei die dreifache Methylierung des Lysinrestes an Position 27 des Histons 3 genannt, diese kann zur Unterdrückung der Genexpression führen.

Definition 2 (Euchromatin vs. Heterochromatin) Im Interphasekern liegt das meiste Chromatin ausgestreckt, und undeutlich anfärbbar vor. Es handelt sich um das **Euchromatin**, das typischerweise eine hohe Gendichte und Transkriptionsaktivität hat. Im Gegensatz dazu steht das **Heterochromatin**, ein Teil des Chromatins, das stets in einem hoch kondensierten Zustand bleibt und dunkel färbbbare Bereiche bildet. Das Heterochromatin wird in *Konstitutives* und *fakultatives* Heterochromatin unterteilt.□

Das **Nukleosom** stellt die Grundeinheit der Verpackung des Chromatins dar und besteht aus einem Kern mit 8 Histoneinheiten. Benachbarte Nukleosomen werden über einen kurzen DNA-Abschnitt (spacer) miteinander verbunden und zeigen somt eine Perlenschnur-Struktur.

Die Chromosomen weisen verschiedene Strukturelemente auf.

Definition 3 (Zentromer) Jedes Chromosom besitzt ein einziges Zentromer. Es handelt sich um den Bereich der primären Einschnürungsstelle eines Metaphase-Chromosoms, an dem die beiden Chromatiden zusammenhängen. **Schwesterchromatiden** sind identische Kopien eines Chromosoms, die über das Zentromer miteinander verbunden sind (und die während der

Mitose bzw. Meiose voneinander getrennt werden). Menschliche Zentromere bestehen zum Teil aus α-satelliten-DNA: Tandemartig wiederholte DNA-Sequenzen ($n \times 171$ bp = 200–9.000 kb). Das **Kinetochor** besteht aus Proteinstrukturen, welche Zentromer-Spindel-Assoziationen und Chromosomenbewegungen vermitteln □

Definition 4 (Telomer) Das Telomer ist die Abschluss-Struktur der eukaryontischen Chromosomen und hat mehrere Funktionen wie die Aufrechterhaltung der Struktur des Chromosoms, die Sicherung einer vollständigen Replikation sowie die Chromosomenpositionierung Beim Menschen finden sich an den Telomeren $(TTAGGG)_n$-Wiederholungen über 3-20 kb, vor denen noch 100–300 kb telomerassoziierte Wiederholungen liegen □

Auf die Inhalte der Chromosomen, die Gene, die regulatorischen Elemente, den "Junk", werden wir in späteren Vorlesungen zu sprechen kommen. Im Folgenden soll kurz auf das Erscheinungsbild und die Klassifikation der menschlichen Chromosomen eingegangen werden.

Definition 5 (p-Arme und q-Arme) Die Chromatiden werden durch das Zentromer in zwei "Arme" unterteilt. Nach der Konvention werden die Chromosomen vertikal dargestellt, wobei der kurze Arm nach oben zeigt und der lange Arm nach unten. Der kurze Arm wird als der p-Arm bezeichnet (p für petite), der lange als der q-Arm (q als der nächste Buchstabe nach p). Je nach Lage des Centromers spricht man von **metazentrischen** (Centromer in der Mitte), **akrozentrischen** (am Ende, der kürzere Arm sehr klein) oder **submetazentrischen** (zwischen Mitte und Ende) Chromosomen. □

Nach der oben aufgeführten Definition werden die menschlichen Chromosomen in folgende Gruppen eingeteilt:

Group A (1-3)	Große metazentrische Chromosomen
Group B (4-5)	Große submetazentrische Chromosomen
Group C (6-12, X)	Mittel-große metazentrische oder submetazentrische Chromosomen
Group D (13-15)	Mittel-große akrozentrische Chromosomen mit Satelliten
Group E (16-18)	Relativ kleine metazentrische oder submetazentrische Chromosomen
Group F (19-20)	Kleine metazentrische Chromosomen
Group G (21,22,Y)	Kleine akrozentrische Chromosomen mit Satelliten (Y hat keine Satelliten)

Im *International System for Cytogenetic Nomenclature* (ISCN) wurde ein System festgelegt, womit die chromosomalen Regionen zu bezeichnen sind. Die Regionen auf dem p- und q-Arm werden als p1, p2, p3,..., q1, q2, q3 von zentromer nach telomer gezählt. Diese Regionen werden wiederum in **Banden** , z.B 12p11, 12p12 und **Subbanden** 12p11.1 eingeteilt. Die Qualität der Chromosomenpräparation (Auflösung) bestimmt die Sichtbarkeit der Banden. Die Chromosomenformel oder **Karyotyp** ist eine Diagnose über die Chromosomenstruktur. Zum Beispiel:

48,XY,t(9;22)(q34;q11),+der(22)t(9;22),+mar

- 48: Chromosomenzahl
- XY: Geschlechtschromosomen
- t(9;22)(q34;q11),+der(22)t(9;22): Aberrationen
- +mar: Markerchromosomen

Es folgende mehere (einfachere) Beispiele:

- Ein Karyotyp ist eine verkürzte Auswertung eines Karyogramms, z.B.
  - Normal weiblich: 46,XX
  - Normal männlich 46,XY
  - Trisomie 21, männlich 47,XY,+21

- Turner-Syndrom: 45,X
- Klinefelter-Syndrom: 47,XXY
- Translokation: 46,XX,t(4,8)(p16;p23)
- Deletion: 46,XX,del(2)(q23q32)

Definition 6 (Chromosomenanomalien) Eine Chromosomenanomalie bezeichnet eine mikroskopisch sichtbare strukturelle Abnormalität in einem oder mehreren Chromosomen oder aber eine atypische Anzahl von Chromosomen. Die häufigste Ursache der Chromosomenanomalien liegt in einem Fehler bei der Meiose oder Mitose. □

Im folgenden beschreiben wir die wichtigsten Chromosomenanomalien.

Definition 7 (Nummerische Chromosomenanomalien) Der Begriff **Nummerische Chromosomenanomalie** beschreibt das Vorliegen einer abnormalen Zahl von sonst normalen Chromosomen. Zum Beispiel ist das zusätzliche Vorliegen eines einzelnen Chromosoms eine **Trisomie** und das Fehlen eines einzelnen Chromosoms eine **Monosomie**. Das Down-Syndrom beruht auf einer Trisomie 21. Die Trisomie 21 sowie Trisomie 18 und Trisome 13 stellen die einzigen mit dem Leben vereinbaren Trisomien von Autosomen dar. Zusätzlich vorliegende Geschlechtschromosomen (**Gonosomen**) resultieren hingegen meist in milde Phänotypen. Die einzige mit dem Leben vereinbare Monosomie ist die Monosomie X (Turnr-Syndrom). Alle anderen Monosomien und Trisomien sind nur als **Mosaike** mit dem Leben vereinbar, d.h. dass die Monosomie bzw. Trisomie nur in einem Teil der Zellen vorliegt. □

Definition 8 (Translokation) Eine **Translokation** ist ein Stückaustausch zwischen zwei oder selten mehreren Chromosomen. **Balancierte Translokationen** sind Translokation ohne Verlust oder Zugewinn genetischen Materials. Balancierte Translokationen sind meist familiär und gehen in der Regel ohne klinische Auffälligkeiten einher. Es besteht jedoch häufig ein erhöhtes Risiko, die Translokation in unbalancierter Form an Kinder zu vererben, da während der Meiose die Paarungsverhältnisse der homologen Chromosomen durch die Translokation gestört werden. Keimzellen mit unbalancierten Translokationen können die Fehlgeburten oder Fehlbildungen bei den betroffenen Kindern führen. □

Definition 9 (Robertson-Translokation) Eine **Robertson-Translokation** resultiert von der Fusion der langen Arme zweier akrozentrischen Chromosomen (13,14,15,21,22) unter Verlust der kurzen Arme beider Chromosomen. Durch den Verlust der kurzen Arme verringert sich zwar die Anzahl der Chromosomen auf 45, jedoch enthalten die kurzen Arme dieser Chromosomen lediglich Gene für rRNA, die sonst im Genom in mehrfachen Kopien vorhanden sind. Es besteht jedoch wie bei balancierten Translokationen ein Risiko einer unbalancierter Weitergabe der Translokationschromosomen. □

Definition 10 (Inversion) Die **Inversion** (oder Rotation eines chromosomalen Segmentes um 180° innerhalb eines Chromosoms) erfolgen meinst ohne Verlust oder Zugewinn genetischen Materials und gehen nicht mit klinischen Auffälligkeiten einher. Es besteht in bestimmten Konstellationen ein erhöhtes Risiko, in der Meiose unbalancierte, rekombinierte Chromosomen zu generieren und diese an Kinder zu vererben. □

Definition 11 (Duplikation) Eine **Duplikation** bedeutet eine Verdopplung eines chromosomalen Segmentes. Dies ist meist mit klinischen Auffälligkeiten verbunden. □

Definition 12 (Deletionen) Eine **Deletion** bedeutet den Verlust eines eines chromosomalen Segmentes. Dies ist meist mit klinischen Auffälligkeiten verbunden. □

Definition 13 (Mikrodeletion) Eine **Mikrodeletion** ist eine kleine Deletion (< 5 Mb), die nicht durch zytogenetische Routineanalysen sichtbar ist. □

Definition 14 (Markerchromosom) Ein zusätzlich vorliegendes Chromosom(fragment), dessen chromosomale Zusammensetzung mit den verwendeten Methoden nicht bestimmt werden konnte. □

1.1 ”Genomkrankheiten” (Engl. genomic disorders)

Mutationen einzelner Basenpaare (z.B. Missense-Mutationen, Nonsense-Mutationen, vgl. Vorlesung 10–11) stellen seit Jahrzehnten bekannte Mechanismen hereditärer Erkrankungen dar. Innerhalb der letzten Jahre (und deshalb in der 2005 veröffentlichen Ausgabe des Lehrbuchs von Strachan und Read kaum erwähnt) wurde dank verbesserter diagnostischer Technologien wie Array-CGH die wichtige Rolle von größeren, zwischen Hunderten und Tausenden von Basenpaaren umfassenden Umbauten der Chromosomen erkannt, die jedoch in der Regel zu klein sind, um mit der konventionellen Zytogenetik detektiert zu werden. Strukturelle Besonderheiten, oder die ”Architektur” des menschlichen Genoms führen zu einer regionell unterschiedlichen Anfälligkeit für solche chromosomale Umbauten bzw. für genomische Instabilität. Aus diesem Grund hat James Lupski den Begriff **genomic disorders** (”Genomkrankheiten”) für diese Gruppe von Krankheiten eingeführt[1]

Man hat zahlreiche nahezu identische chromosomale Umbauten bei unverwandten Personen beobachtet, d.h., rezidivierende Rearrangements. Die meisten Umbauten vopn dieser Sorte werden durch nicht allelische homologe Rekombination (NAHR) zwischen zwei LCRs verursacht.

Definition 15 (LCR) **Low-copy repeats** bezeichnen regionspezifische DNA-Blöcke von 10–200 Kilobasen, die eine Ähnlichkeit untereinander von 95% bis 97% aufweisen. Es handelt sich hierbei um duplizierte Sequenzen, welche 5-10% des menschlichen Genoms ausmachen. Sie können Gene, Pseudogene, endogene virale Sequenzen enthalten und kommen bevorzugt in Zentromer-/Telomernähe vor. □

Aufgrund des hohen Grades an Sequenzähnlichkeit zwischen nichtallelischen Kopien eines LCRs kann es bei der Meiose oder Mitose dazu kommen, dass nicht die allelische Kopien sondern nichtallelische Kopien eines LCRs aneinanderheften. Die Folge einer Rekombination (Crossing-over) in dieser Kostellation kann ein genomischer Umbau sein (Fig. 1.1).

Solche Umbauten sind eine relativ häufige Ursache hereditärer Krankheit. Zum Beispiel wird die neurologische Erkrankung Charcot-Marie-Tooth-Krankheit Typ 1A (CMT1A) durch eine heterozygote Duplikation eines 1,4 Mb großen Abschnitts auf Chromosom 17 verursacht, was dazu führt, dass ein zusätzliches Exemplar des Gens PMP22 vorliegt. Dieses Gen kodiert für das periphere Myelinprotein. Die hereditäre Neuropathie mit Anfälligkeit für Druckparesen (HNPP) auf der anderen Seite wird durch eine heterozygote Deletion genau desselben Abschnitts verursacht. Offensichtlich sind die peripheren Nerven auf die ”richtige Dosis” der PMP22-Genaktivität angewiesen.

[1] s. Lupski JR, Stankiewicz P (2005) Genomic Disorders: Molecular Mechanisms for Rearrangements and Conveyed Phenotypes. PLoS Genet 1(6): e49. doi:10.1371/journal.pgen.0010049.

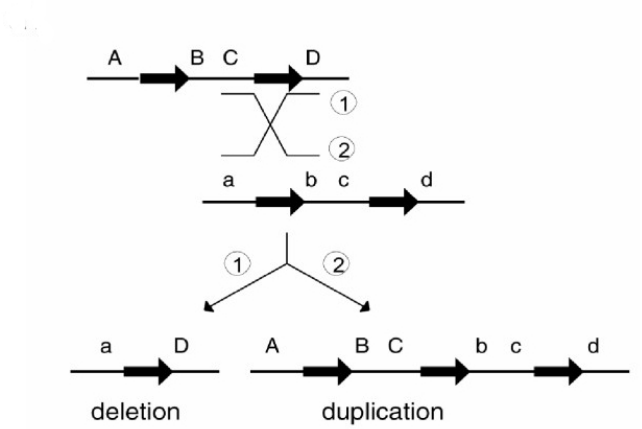


Abbildung 1.1: Nicht allelische homologe Rekombination (NAHR). Die nicht allelischen Kopien eines LCRs begünstigen eine homologe Rekombination. Sind die zwei beteiligten LCRs auf dem selben Chromosom und in derselben Orientierung gelegen, führt die NAHR zur Duplikation und/oder Deletion. Abbilbung entnommen aus Gu W, Zhang F, Lupski JR (2008) Mechanisms for human genomic rearrangements PathoGenetics 2008, 1:4.

Kapitel 2

Populationsgenetik

2.1 Einführung

1 Definition Die **Populationsgenetik** befasst sich mit genetischen Strukturen in Populationen und ihren Veränderungen im Laufe der Evolution.

Die Populationsgenetik ist eine wichtige Grundlage der genetischen Epidemiologie. Wichtige Begriffe in der Populationsgnetik sind:

- **Transmission**: Vererbung von einer Generation zur nächsten
- **Selektion**: genetische Auslese
- **Mutation**: Neuentstehung genetischer Varianten

Das Ziel der nächsten sechs Vorlesungen ist es, ein Verständnis der biologischen und mathematischen Grundlagen der Populationsgenetik zu vermitteln. Kopplungsanalysen (Anwendung: bislang unbekannte Krankheitsgene zu entdecken) und Sequenzkonservierung (Grundlage: Selektionsdruck gegen den Hintergrund neutraler Veränderungen aufgrund von genetischem Drift, Anwendung z.B. Sequenzalignment) sind wichtige Beispiele von bioinformatischen Themen, die ein Grundverständnis der Populationsgenetik voraussetzen.

Es werden kurze matlab Skripten zur Verdeutlichung der Inhalte angeboten. Die Skripten können von `http://compbio.charite.de` heruntergeladen werden. Fragen zu matlab und den Skripten können in den Übungen besprochen werden. matlab selber ist kein Prüfungsgegenstand, sondern die besprochenen Verfahren und Algorithmen, die mit den Skripten verdeutlicht werden sollen. Die Skripten können größtenteils auch in octave[1] (open-source und kostenlose matlab-ähnliche Skriptumgebung) verwendet werden (ich habe die Skripten mit Octave 3.0 unter Debian-Linux getestet).

2.2 Das Wachstum von Populationen

Bakterien vermehren sich durch Zweiteilung. Das Darmbakterium *Escherichia coli* wächst bis zu einer Dichte von $2-3\times10^9$ Zellen/ml heran. Bei dieser Konzentration hemmen Ausscheidungsprodukte die weitere Vermehrung. Man kann die Wachstumskinetik von Bakterien in vier Phasen unterteilen (vgl. Abb. 2.1):

- die ***lag*–Phase** (A). Es handelt sich um eine Phase, in der die Bakterien sich erst an die neuen Kulturbedingungen adaptieren und den Stoffwechsel auf Teilung einstellen. ('lag'=Verzögerung)
- die **logarithmische Wachstumsphase** (B). Die Anzahl der neu entstehenden Individuen ist proportional zur Anzahl der bestehenden Individuen.
- die **stationäre Phase** (C). Die Nähstoffe im Kulturmedium werden weniger bzw. toxische Ausscheidungsprodukte häufen sich an.
- eine **Abbauphase** (D) Die Nähstoffe sind verbracht und können den Energiebedarf der Bakterien nicht mehr decken, es kommt zum Tod der Bakterien.

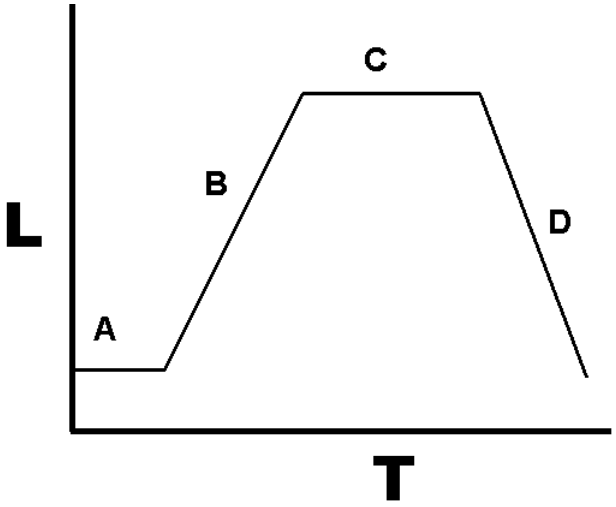


Abbildung 2.1: Das Wachstum von Bakterien in Kultur kann in vier Phasen unterteilt werden: lag-Phase (A), logarithmische Wachstumsphase (B), stationäre Phase (C) und Abbauphase (D). Gezeigt ist $\boldsymbol{L} = \log(N)$, wo N die Anzahl von Individuen angibt und $\boldsymbol{T}$, Zeit. Bildquelle: Wikipedia Commons.

Im folgenden wollen wir ein einfaches Modell für das Wachstum von *E. coli* Zellen beschreiben, das zumindest die logarithmische und die stationäre Phase beschrieben kann. Mit einem ähnlichen Modell kann auch die Konkurrenz zwischen zwei unterschiedlichen Organismen modelliert werden (s. weiter unten).

Betrachte eine Population mit x Individuen. Sei $r\Delta t$ die Wahrscheinlichkeit, dass ein einzelnes Bakterium in der Zeit Δt sich teilt. Die Zunahme in der Anzahl der Bakterien ist dann

$$\Delta x = x \cdot r\Delta t \tag{2.1}$$

Für $\Delta t \to 0$ erhält man die Differentialgleichung:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = rx \tag{2.2}$$

Die Lösung ist

$$x = x_0 e^{rt} \tag{2.3}$$

wobei $x_0 = x(0)$, d.h., die Anzahl der Individuen zu Zeit $t = 0$. Indem wir beide Seit logarithmisieren, erhalten wir

$$\log x = \log x_0 + rt \tag{2.4}$$

Wir wollen nun ein erstes matlab Skript entwickeln, um die logarithmische Wachstumsphase darzustellen.

```
x0=1.0;
r=0.01;
t=linspace(0,200,10); % 10 linear verteilte Punkte zw. 0 und 200
lnx = log(x0)+r*t;    % Vektoroperation, alles auf einmal!
plot(t,lnx,'ro-');    % plot x,y in rot ('r'), Datenpunkt als Kreise ('o')
                      % verbunden mit Linie ('-')
axis([0 300 0 3]);    % Achsen einstellen [xmin xmax,ymin ymax]
xlabel('Zeit');       % Beschriftung der X-Achse
ylabel('Wachstum')
title('Logarithmische Wachstumsphase');
```

Die Kommandos können eins nach dem anderen in matlab eingegeben werden oder als Skript in einer Datei (z.B. logphase.m) gespeichert werden und durch eingabe von `>logphase` ausgeführt werden. Das Ergebnis ist in Abb. 2.2 zu sehen.

Das Wachstum kann jedoch nicht ewig logarithmisch weitergehen, da früher oder später die Resourcen verbraucht werden. Die Auswirkung der Resourcenbegrenzung wird mittels einer logistischen[2] Gleichung beschrieben. Intuitiv kann man sagen,

[1] s. http://www.gnu.org/software/octave/

[2] Logistisches Wachstum beschreibt einen Wachstumsprozess mit Selbstbegrenzung.

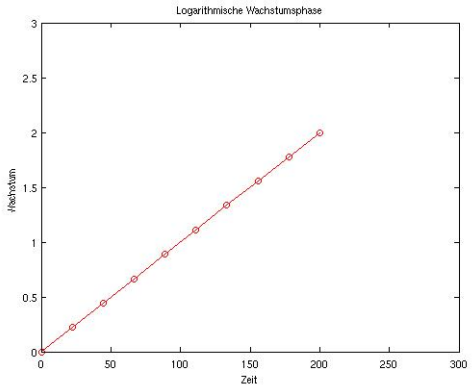


Abbildung 2.2: Logarithmische Wachstumsphase

dass sich die Population einer Größe, nähert, die längerfristig stabil bleiben kann (die Konstante `K` gibt die *Kapazität*, englisch 'carrying capacity', an.):

$$\frac{\mathrm{d}x}{\mathrm{d}t} = rx(1 - x/K) \tag{2.5}$$

Ist $x \ll K$ am Anfang des Wachstums, dann gilt $\frac{\mathrm{d}x}{\mathrm{d}t} \approx rx$, was im Einklang mit der Beobachtung steht, dass sich Populationen am Anfang i.d.R. exponentiell vermehren. Ist $x < K$, dann ist $\frac{\mathrm{d}x}{\mathrm{d}t} > 0$, d.h., die Population wächst. Ist $x > K$, d.h., die aktuelle Population liegt oberhalb der Kapazität, dann ist $\frac{\mathrm{d}x}{\mathrm{d}t} < 0$ und die Population verringert sich.

Die Lösung der logistischen Gleichung (die Herleitung findet sich im Anhang) lautet:

$$x = \frac{Ce^{rt}}{1 + Ce^{rt}/K} \tag{2.6}$$

Die Konstante C lässt sich anhand des Wertes x_0 und der Kapazität K berechnen als

$$C = \frac{Kx_0}{K - x_0}$$

Wir können die Ergebnisse wie folgt plotten.

```
x0=0.1;
r=0.01;
K=1.5;

t=linspace(0,750,100); % 100 linear verteilte Punkte zw. 0 und 750
C=(K*x0)/(K-x0);
f=@(t) C.*exp(r.*t) ./ (1 + C.*exp(r.*t)./K);
x=f(t);                % Vektoroperation, alles auf einmal!
plot(t,x,'ro-');       % plot x,y in rot ('r'), Datenpunkt als Kreise ('o')
                       % verbunden mit Linie ('-')
axis([0 750 0 2]);     % Achsen einstellen [xmin xmax,ymin ymax]
xlabel('Zeit')         % Beschriftung der X-Achse
ylabel('Wachstum')     % Beschriftung der Y-Achse
title('Logistisches Wachstum');
```

Die Zeile `f=@(t) C.*exp(r.*t) ./ (1 + C.*exp(r.*t)./K);` weist der Variablen f eine anonyme Funktion zu, welche die logistische Gleichung berechnet. Die Schreibweise `./` und `.*` bedeutet elementweise Division und Multiplikation, so dass man mit der einzigen Zeile `x=f(t)` die x-Werte für den gesamten Vektor t auf einmal berechnet (Abb. 2.3). Das Skript kann von der Kurs-Website heruntergeladen werden (`logistic.m`).

Aufgabe: Verändere das Skript und beobachte das Verhalten des Systems mit unterschiedlichen Parametern. Was passiert, wenn x0 > K? Was bestimmt, wie schnell x seinen Gleichgewichtswert annimmt (bzw. sich ihm annähert)?

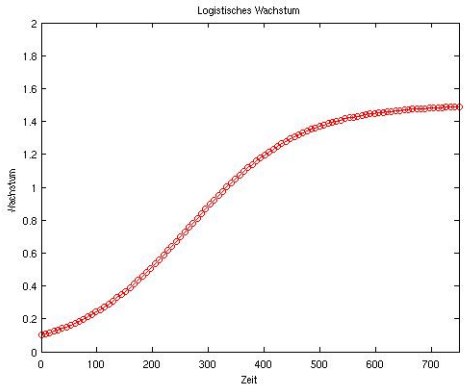


Abbildung 2.3: Logistische Wachstumsphase

2.3 Konkurrenz

Unter welchen Umständen wird ein Organismus einen anderen von einer gegebenen Nische verdrängen? Eine Antwort auf diese Frage gibt eine veränderte Fassung von Gleichung (2.5). Seien x und y jeweils die Anzahl von Bakterien vom Typ X und Y. Das Wachstum von X und Y kann beschrieben werden durch die Differentialgleichungen

$$\frac{\mathrm{d}x}{\mathrm{d}t} = r_1 x(1 - x/K_1) \tag{2.7}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = r_2 y(1 - x/K_2) \tag{2.8}$$

Da die Gleichungen nicht miteinander gekoppelt sind, wird wie im obigen Beispiel x bis zur Kapazität K_1 anwachsen und y bis zur Kapazität K_2. Eine solche Situation wäre dann gegeben, wenn unterschiedliche Resourcen jeweils das Wachstum von X und Y begrenzten, d.h., wenn die beiden Organismen nicht in Konkurrenz zueinander ständen. Nehmen wir andererseits an, dass das Wachstum von X auch durch das Wachstum von Y begrenzt wird und umgekehrt, dann haben wie die gekoppelten Differentialgleichungen

$$\frac{\mathrm{d}x}{\mathrm{d}t} = r_1 x(1 - (x+y)/K_1) \tag{2.9}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = r_2 y(1 - (x+y)/K_2) \tag{2.10}$$

Falls $K_1 \neq K_2$, wird der eine Organismus den anderen komplett ersetzen. Falls $K_1 > K_2$ wird der Wert von x ansteigen, bis $x+y = K_1$. Sobald $x+y > K_2$ (vgl. 2.10) gilt $\frac{\mathrm{d}y}{\mathrm{d}t} < 0$, d.h., der Wert von y nimmt ab (vgl. Fig. 2.4). Für Neugierige haben wir matlab-Code zur Lösung dieses Systems im Anhang 5.2 erklärt.

2.4 Mutationen: Terminologie

Definition 16 (Locus) Unter **Locus** (Latein: "Ort") versteht man in der Genetik ein Gen, ein Nukleotid oder einen Sequenzabschnitt (die jeweilige Bedeutung wird aus dem Kontext klar). □

Definition 17 (Mutationsrate) Die **Mutationsrate** μ ist definiert als die Anzahl neuer Mutationen pro Generation pro Gamet bezogen auf einen Locus. Beispiele für Mutationsraten sind:

- Gene: in der Größenordnung von 10^{-4} bis 10^{-6}

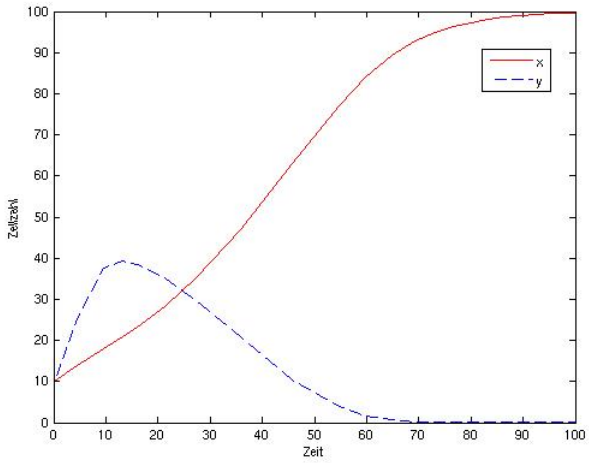


Abbildung 2.4: Konkurrenz zwischen zwei Bakterienstämme. $r_1 = 0.1$, $r_2 = 0.4$, $K_1 = 100$, $K_2 = 60$. Sowohl x und y beginnen mit jeweils 10 Individuen. y hat eine höhere Wachstumsrate als x aber eine geringere Kapazität.

- Mikrosatelliten: 10^{-2} bis 10^{-4}
- SNPs: 10^{-8} bis 10^{-9} □

In einer späteren Vorlesung werden wir auf die Bedeutung der einzelnen Mutationsklassen (Missense, Nonsense, Repeatexpansion, ...) zurückkommen.

Definition 18 (Allel) Ein **Allel** bezeichnet eine mögliche Variante eines Locus (etwa das ein Nukleotid an einer bestimmten Position ein 'A' oder 'C' sein kann; dann liegen zwei Allele vor). □

2.5 Genetische Drift

Das im Abschnitt 2.3 beschriebenen Modelle ist deterministisch. In (ausreichend) großen Populationen trifft diese Annahme annähernd zu. Insbesondere bei kleinen Populationen kommt es zu zufälligen Veränderungen der Verteilungen der Allele. Solche Fluktuationen werden als genetische Drift (englisch, *drift*=sich treiben lassen). Die Folge ist, dass es ohne das Wirken anderer Mechanismen (insbesondere ohne das Wirken der "natürlichen Auslese") zu Schwankungen bei den Allel- und Genotyphäufigkeiten kommt.

Das folgende Modell beschreibt die genetische Drift in einer sich asexuell fortpflanzenden Population von N Individuen mit zwei Allelen an einem Locus, a und A. Die Anteile von a und A in der Population werden von p und q angegeben, d.h., $pN + qN = N$. Seien p' und q' die Anteile von a und A in der nächsten Generation. Falls jedes Individuum in der Population genau einen Nachkommen hat, das gilt $p = p'$ und $q = q'$.

Ein realistischeres Modell ist aber folgendes. Jedes der N Nachkommen in jeder Generation wird zufällig einem der N Eltern zugewiesen. Dann gilt eine Wahrscheinlichkeit von p für jedes Nachkommen, vom Typ a zu sein bzw. q vom Typ A zu sein[3]. Wir können die genetische Drift mit folgendem matlab-Skript simulieren:

```
N=100; % Populationsgroesse
p=0.5; % Anteil Typ A in Population
```

[3] Die Wahrscheinlichkeit, dass es dann x Individuen vom Typ A in der nächsten Generation folgt der Binomialverteilung

$$P(x) = \binom{N}{x} p^x q^{N-x} = \frac{N!}{(N-x)!x!} p^x q^{N-x} \tag{2.11}$$

, wobei wir denselben Effekt durch den relativ einfachen matlab/octave-Skript erreichen.

```
ngenerations = 200;

pop=rand(N,1)>0.5;

A=zeros(ngenerations,1);% Vektor mit Anzahl von Allel A Individuen

for i=1:ngenerations
    A(i) = sum(pop);
    ind=ceil(N*rand(N,1));
    pop=pop(ind);
end

% Ergebnis plotten
clf; % clear current figure.
gen = [1:ngenerations]';
plot(gen,A,'ro-');

axis([ 0 ngenerations 0 100]); % Achsen einstellen [xmin xmax ymin ymax]
xlabel('Generation');       % Beschriftung der X-Achse
ylabel('Anteil der Individuen mit Allel A');
title('Genetische Drift');
```

Die Skript `gdrift.m` kann von `http://compbio.charite.de/index.php/genetikfuerbioinformatiker.html` heruntergeladen werden. Eine Erklärung des Skriptes findet sich im Anhang 5.3. Das Ergebnis eines typischen Durchlaufs wird in Abb. 2.5 dargestellt.

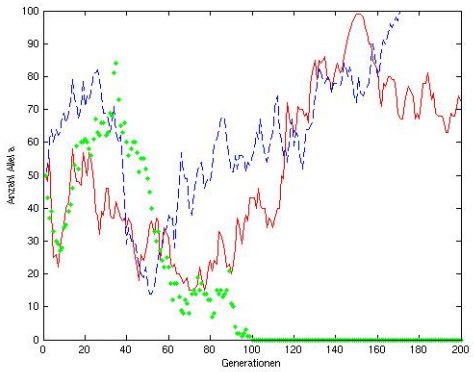


Abbildung 2.5: Genetische Drift. Drei Simulationen sind dargestellt.

Frage: Was ist der Einfluss des Parameters *N* (Populationsgröße) auf die genetische Drift?

Genetische Drift: Zusammenfassung

1. In endlichen Populationen schwanken die Häufigkeit der Allele durch reinen Zufall, ohne das Wirken der natürlichen Auslese.
2. Je kleiner die Population, desto ausgeprägter sind die Schwankungen
3. In einer endlichen Population kommt es mit einer Wahrscheinlichkeit von 100% zur Fixierung eines Allels (Dies gilt nur für idealisierte Populationen, bei denen z.B. keine neuen Mutationen auftreten, wie in unserer Simulation).

Kapitel 3

Hardy-Weinberg-Gesetz

In diesem Kapitel sollen Allelverteilungen in großen Populationen eines diploiden Organismus untersucht werden.

Definition 19 (Genotyp) Der **Genotyp** eines Individuums beschreibt seine Ausstattung von Genen bzw. Allelen. Das Wort Genotyp kann auch verwendet werden, um nur ein Genlocus zu beschreiben. Hat ein Gen z.B. zwei Allele, A und a, dann gibt es die Genotypen AA, Aa und aa. □

Das **Hardy-Weinberg-Gesetz** beschreibt die Verteilung von Allelen in großen Populationen. Wir betrachten ein Locus mit zwei Allelen A und a mit entsprechenden Häufigkeiten p und $q = 1 - p$. Das Hardy-Weinberg-Gesetz stellt eine Beziehung zwischen den Allelhäufigkeiten und den Genotyphäufigkeiten her. Unter bestimmten Voraussetzungen können wir die Häufigkeiten der Genotypen AA, Aa und aa als p^2, $2pq$, und q^2 berechnen. Dies ist eine einfache Anwendung der Binomialverteilung (Tabelle 3.1):

	A	a
A	p^2	pq
a	pq	q^2

Tabelle 3.1: Die väterlichen Allele werden in der obersten Zeilen gezeigt, die mütterlichen in der linken Spalte. Die Tabelle zeigt die Häufigkeiten der entsprechenden Allelkombinationen, d.h., Genotypen.

Das Hardy-Weinberg-Gesetz beschreibt die Verteilung der Genotypen in einer idealen Population. Abweichungen vom Hardy-Weinberg-Gleichgewicht können vorliegen, wenn

- Keine Zufallspaarung existiert. Dies kann der Fall sein, wenn unterschiedliche Allele einen Einfluss auf äußerliche Merkmalen wie Körpergröße und somit ggf. auf die Paarungswahrscheinlichkeit haben.
- Eine Selektion für oder gegen ein Allel vorliegt (wir werden diese Situation weiter unten analysieren)
- die Population aus Subpopulation mit unterschiedlichen Allelverteilungen besteht. Dann beobachtet man mehr homozygoten und weniger Heterozygoten als man allein aufgrund der Allelhäufigkeiten erwarten würde (**Wahlund-Effekt**).

Wir können nun zeigen, dass die Allelfrequenzen von Generation zu Generation unter diesen Vortaussetzung stabil bleiben. Definieren wir P als die Häufigkeit von Individuen mit dem Genotyp AA, H als die Häufigkeit von Individen mit dem Genotyp Aa, und Q als die Häufigkeit von Individuen mit dem Genotyp aa. Nehmen wir an, dass eine Zufallspaarung (in Bezug auf die Allele A und a) innerhalb der Population vorliegt, dann können wir die Häufigkeiten der Genotypen und der Allele in der nächsten Generation wie in Tabelle 3 berechnen.

Paarung	Häufigkeit	Nachkommen AA	Aa	aa
$AA \times AA$	P^2	P^2	-	-
$AA \times Aa$	$2PH$	PH	PH	-
$AA \times aa$	$2PQ$	-	$2PQ$	-
$Aa \times Aa$	H^2	$H^2/4$	$H^2/2$	$H^2/4$
$Aa \times aa$	$2HQ$	-	HQ	HQ
$aa \times aa$	Q^2	-	-	Q^2
Gesamt	$(P+H+Q)^2$	$(P+H/2)^2$	$2(P+H/2)(H/2+Q)$	$(H/2+Q)^2$
	$=1$	$=p^2$	$=2pq$	$=q^2$

Tabelle 3.2: **Hardy-Weinberg-Gleichgewicht**. Ist in Generation i die Häufigkeit von Allel A p und diejenige von Allel a q mit $p+q=1$, dann ist die Häufigkeit von AA-Homozygoten $P=p^2$, von Aa-Heterozygoten $H=2pq$ und von aa-Homozygoten $Q=q^2$. Die Häufigkeit von Paarung von Individuen mit diesen drei Genotypen wird in der zweiten Spalte, die Häufigkeiten von Nachkommen mit den drei Genotypen in Generation $i+1$ in den letzten drei Spalten angegeben. Ersichtlich ist, dass die Allelfrequenzen von A und a und die Genotypfrequenzen von AA, Aa und aa sich von Generation zu Generation nicht ändern (falls die Bedingungen des Hardy-Weinberg-Gesetzes zutreffen). Bemerke, dass die Häufigkeit der Paarung $AA \times Aa$ $2PH$ ist, weil es zwei mögliche Fälle gibt: entweder der Vater hat AA und die Mutter Aa oder umgekehrt.

Definition 20 (Gamet) **Gameten**, auch bekannt als Geschlechtszellen oder Keimzellen, sind haploide Zellen, die von sich geschlechtlich fortpflanzenden Organismen erzeugt werden, z.B. sind Eizellen und Spermien Gameten. Eine **Zygote** ist eine diploide Zelle, die durch Verschmelzung zweier Gameten entsteht, z.B. eine befruchtete Eizelle. □

Definition 21 (Fitness) Im populationsgenetischen Kontext ist die **Fitness** ein Maß für die Reproduktionswahrscheinlichkeit der verschiedenen Genotypen. Vom Blickwinkel der Populationsgenetik ist der Grund einer reduzierten Fitness unbedeutend und kann z.B. durch reduzierte Überlebenswahrscheinlichkeit oder reduzierte Fruchtbarkeit bedingt sein. Die Fitness eines bestimmten Genotyps ist die zu erwartende Anzahl von Nachkommen in der nächsten Generation. Die Fitness wird auf ein bestimmtes Stadium im Lebenszyklus bezogen, in der Regel auf den Zygoten. Die Fitness der Zygoten vom Typ A ist dann die zu erwartende Anzahl der Nachkommen dieser Zygoten unter den Zygoten der nächsten Generation. Bemerke, dass sich die Fitness auf Genotypen bezieht, nicht auf die einzelnen Allele. □

2 Beispiel Betrachten wir drei Genotypen `AA`, `AB` und `BB` und beschreiben wir die relative Fitness der Genotypen. Diese ist das Produkt aus der Überlebenswahrscheinlichkeit und (gegeben, dass ein Individuum überlebt) die durchschnittliche Zahl von Gameten, die an die nächste Generation weitergegeben werden.

- `AA` hat eine Überlebenswahrscheinlichkeit von 9/10. Überlebende geben durchschnittlich 5 Gameten an die nächste Generation
- `AB` hat eine Überlebenswahrscheinlichkeit von 8/10. Überlebende geben durchschnittlich 4 Gameten an die nächste Generation
- `BB` hat eine Überlebenswahrscheinlichkeit von 7/10. Überlebende geben durchschnittlich 2 Gameten an die nächste Generation

Die absolute Fitness von `AA` ist dann $9/10 \times 5 = 4,5$, die von `AB` ist $8/10 \times 4 = 3,2$, die von `BB` ist $7/10 \times 2 = 1,4$. In der Regel interessieren wir uns mehr für die relative Fitness. Bezogen auf die Fitness von `AA` ist die relative Fitness von `AB` $3,2/4,5 = 0,71$ und die relative Fitness von `BB` $1,4/4,5 = 0,31$.

Die oben angegebene Definition von Fitness unterliegt folgenden Annahmen:

- Die Fitness im populationsgenetischen Sinne bezieht sich nicht auf ein Individuum sondern auf eine Klasse von Individuen, z.B. alle Individuen mit dem Genotyp AA, die Phrase "zu erwartende Anzahl" von Nachkommen ist also der Durchschnitt für die Gruppe.

- Die Fitness ist immer auf eine bestimmte Umgebung bezogen. Ein Genotyp, der in einer Umgebung einen Vorteil mit sich bringt, kann in einer anderen Umgebung von Nachteil sein.

3.1 Die Ausbreitung eines günstigen Allels in einer Population

In diesem Abschnitt werden wir die Ausbreitung eines vorteilhaften Allels in einer Population untersuchen. Sei A das vorteilhafte Allel und a das nachteilhafte. Sei p_n die Frequenz von A in Generation n. Die Fitness von AA wird angegeben mit $1+s$, die von Aa mit $1+hs$ und die von aa mit 1. Wir können den Effekt eines dominanten, rezessiven oder intermediären Allels mit unterschiedlichen Werten von h simulieren:

- $h = 1$: dominant

- $h = 0$: rezessiv

- $h = 1/2$: intermediär

In der folgenden Simulation nehmen wir an, dass die unterschiedliche Fitness allein durch eine unterschiedliche Überlebenswahrscheinlichkeit bedingt ist, dass aber Überlebende die gleiche Fruchtbarkeit haben. Dann gilt in Generation n:

Genotyp	AA	Aa	aa
Fitness	$1+s$	$1+hs$	1
Frequenz der Zygoten	p_n^2	$2p_nq_n$	q_n^2
Relative Anteil der überlebenden Erwachsenen	$p_n^2(1+s)$	$2p_nq_n(1+hs)$	q_n^2

Tabelle 3.3: Selektion.

Die letzte Reihe gibt die relativen Anteile der Genotypen unter Erwachsenen Individuen. Die Summe dieser Einträge ist $Z = p_n^2(1+s) + 2p_nq_n(1+hs) + q_n^2$. Dies kann weiter vereinfacht werden: $Z = \left[p_n^2 + 2p_nq_n + q_n^2\right] + p_n^2 s + 2hsp_nq_n = 1 + s\left(p_n^2 + 2hp_nq_n\right)$. Da jeder Genotyp jeweils 2 Allele hat, ist die Summe der relativen Anteile der Allele $2Z$.

Die Frequenz des Allels A in der Generation $n+1$ ist gleich der Frequenz unter überlebenden Erwachsenen von Generation n. Da jedes Individuum jeweils 2 Allele hat, und AA-Individuen zwei A-Allele beisteuern und Aa nur eins, haben wir

$$p_{n+1} = \frac{2 \times \text{relativer Anteil der AA-Erwachsenen} + 1 \times \text{relativer Anteil der Aa-Erwachsenen}}{2Z} \tag{3.1}$$

$$p_{n+1} = \frac{p_n^2(1+s) + p_nq_n(1+hs)}{1+s(p_n^2+2hp_nq_n)} \tag{3.2}$$

Wir können nun aus Gleichung 3.2 eine Finite-Differenz-Gleichung erstellen[1]. Sei $\boxed{\delta p_n}$ die Veränderung in p_n über eine Generation, d.h.,

$$\delta p_n = p_{n+1} - p_n = \frac{p_n^2(1+s) + p_nq_n(1+hs)}{1+s(p_n^2+2hp_nq_n)} - p_n \tag{3.3}$$

[1] Eine finite Differenz ist ein mathematischer Ausdruck von der Form $f(x+h) - f(x)$. Dividiert man diese finite Differenz durch h, erhält man den entsprechenden Differenzenquotienten $\frac{f(x+h)-f(x)}{h}$. Solche Quotienten werden bei der numerischen Differentiation für die näherungsweise Berechnung der Ableitung aus gegebenen Funktionswerten verwendet.

Gibt es keinen Selektionsvorteil für Allel A gegenüber Allel a, dann verändert sich die Verteilung der Allele nicht: $\delta p_n = 0$, falls $s = 0$ (wie man anhand von (3.3) einfach bestätigen kann). Mit folgenden Schritten erreichen wir eine Vereinfachung:

$$\begin{aligned}
\delta p_n &= \frac{p_n^2(1+s)+p_n q_n(1+hs)}{1+s(p_n^2+2hp_nq_n)} - p_n \\
&= \frac{p_n^2(1+s)+p_n q_n(1+hs)-p_n\left[1+s\left(p_n^2+2hp_nq_n\right)\right]}{1+s(p_n^2+2hp_nq_n)} \\
&= \frac{p_n^2+p_n^2 s+p_nq_n+p_nq_nhs-p_n-sp_n^3-2hsp_n^2q_n}{1+s(p_n^2+2hp_nq_n)} \\
&= \frac{p_n(p_n+q_n-1)+p_n^2 s+p_nq_nhs-sp_n^3-2hsp_n^2q_n}{1+s(p_n^2+2hp_nq_n)} \\
&= \frac{p_n^2 s+p_nq_nhs-sp_n^3-2hsp_n^2q_n}{1+s(p_n^2+2hp_nq_n)} \\
&= \frac{sp_nq_n[h-2hp_n]+sp_n^2[1-p_n]}{1+s(p_n^2+2hp_nq_n)} \\
&= \frac{sp_nq_n[h-2hp_n]+sp_n^2[q_n]}{1+s(p_n^2+2hp_nq_n)} \\
&= \frac{sp_nq_n[p_n+h(1-2p_n)]}{1+s(p_n^2+2hp_nq_n)}
\end{aligned}$$

Unter der Annahme, dass der Selektionsvorteil s relativ gering ist, gilt $1+s\left(p_n^2+2hp_nq_n\right) \approx 1$. Wenn der Selektionsvorteil gering ist, dann ist auch die Veränderung pro Generation, und wir können δp_n durch eine Differentialgleichung ersetzen:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = spq[p+h(1-2p)]$$

Interessieren wir uns hauptsächlich für den initialen Anstieg eines neuen vorteilhaften Allels `A`, dann ist p_0 sehr klein and $q_0 \approx 1$. Es folgt:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = sp[p+h]$$

Ist das Allel `A` dominant ($h = 1$), dann gilt $p+h = p+1 \approx 1$. Dann $\frac{\mathrm{d}p}{\mathrm{d}t} = sp$, or $p = p_0 e^s$. Ist das Allel `A` rezessiv ($h = 0$), dann haben wir $p+h = p$ und $\frac{\mathrm{d}p}{\mathrm{d}t} = sp^2$.

Wir können den Anstieg der Frequenz eines vorteilhaften Allels mit folgendem matlab-Code simulieren (die Funktion setzt Gleichung 3.2 um):

```
function A = select(p,s,h,ngenerations)

%The following function calculates p_{n+1} given p_n
f=@(p,q,h,s) (p^2*(1+s) + p*q*(1+h*s))/(1+s*(p^2 + 2*h*p*q));

A = [p ];

for i=2:ngenerations
    q=1-p;
    p = f(p,q,h,s);
    A = [ A; p  ];
end
```

Der Code definiert eine anonyme Funktion, die aus p_n,q_n, h und s p_{n+1} berechnet. Der Vektor A wird mit p_0 initialisiert. In jedem Durchlauf der For-Schleife wird ein neuer Wert für p berechnet und ans Ende von A angefügt. Nach Beendigung der For-Schleife wird A zurückgegeben. Die Simulation wurde dann mit folgenden Kommandos durchgeführt:

```
s=0.1;                  % Selektionsvorteil
ngenerations = 1200;    % Anzahl der Generationen fuer die Simulation
t= [1:ngenerations]';   % Zeitpunkte fuers Plotten

p=0.01;                 % Anfaengliche Frequenz von A

```

```
7
8  h=0; % rezessiv
9  A=select(p,s,h,ngenerations);
10 h=1; % dominant
11 B=select(p,s,h,ngenerations);
12 h=0.5; % intermediaer
13 C=select(p,s,h,ngenerations);
14
15 clf;
16
17 plot(t,A,'r-');
18 hold on;
19 plot(t,B,'b.');
20 plot(t,C,'g--');
21 axis([0 ngenerations 0 1.05]);
```

Mit diesem Code wurde Abb. 3.1 erzeugt.

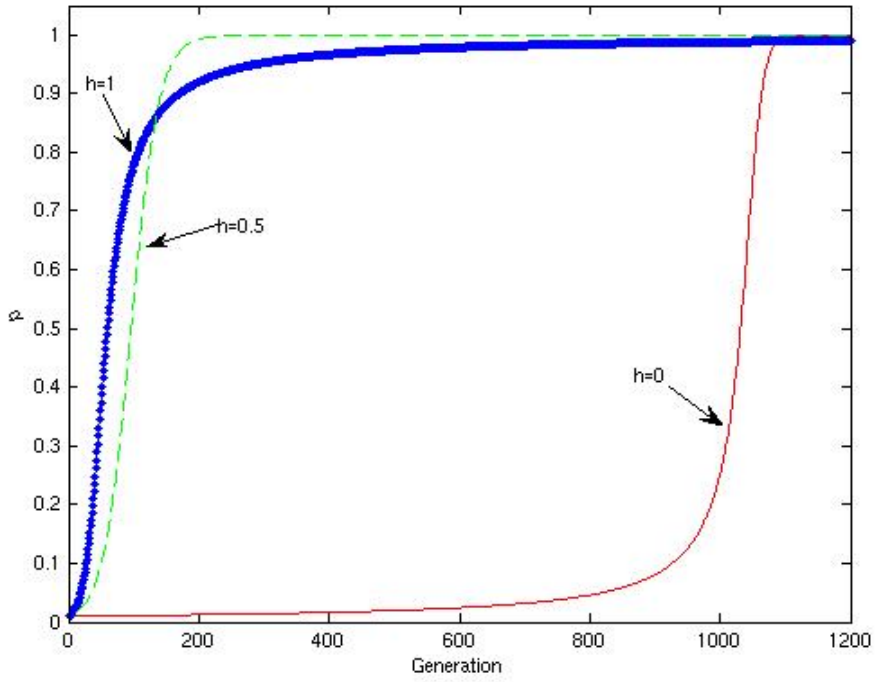


Abbildung 3.1: Selektion. Die Geschwindigkeit, womit sich ein vorteilhaftes Allel in der Population ausbreitet, hängt davon ab, ob sich das Allel dominant ($h = 1$), rezessiv ($h = 0$) oder intermediär (z.B. $h = 0.5$) auswirkt.

3.2 Relevanz für die Humangenetik

Wir werden in einer späteren Vorlesung auf die verschiedenen Erbgänge im Detail eingehen. Für den Moment wollen wir rezessiv und dominante Erkrankungen definieren.

Definition 22 (Autosomal) autosomal (im Gegensatz zu gonosomal) bezieht sich auf Genloci, die nicht auf einem der beiden Geschlechtschromosomen X und Y liegen. Beim Menschen sind die Chromosomen 1–22 Autosomen, die Chromosomen X und Y die Gonosomen (Geschlechtschromosomen). □

Definition 23 (Autosomal dominant) Autosomal dominante Erkrankungen werden durch eine Mutationen in einem autosomalen Gen hervorgerufen. Eine Mutation in einem der beiden Genexemplaren (d.h., auf dem mütterlichen oder auf dem väterlichen Chromosom) reicht zur Merkmalsausprägung (Krankheit). □

Definition 24 (Autosomal rezessiv) Autosomal rezessive Erkrankungen werden durch eine Mutationen in einem autosomalen Gen hervorgerufen. Erst bei Mutation in beiden Genexemplaren kommt es zur klinischen Erkrankung. Heterozygote Mutationsträger sind gesund. □

Bestimmte Mutationen haben so schwerwiegende Folgen, dass Reproduktion für die betroffenen Personen unmöglich ist. Viele Chromosomenstörungen (z.B. Trisomie 13) und auch Mutationen einzelner Gene haben diesen Effekt. Die meisten autosomal dominanten Erkrankungen reduzieren die durchschnittliche Reproduktion der Mutationsträger. Solche Mutationen würden von der Population verschwinden, wenn nicht andere Faktoren, wie Neumutationen, eine Rolle spielten. Daher kann man ein Gleichgewicht zwischen Reproduktionsnachteil und Neumutationsrate annehmen. Sei p die Häufigkeit des mutierten, dominanten Allels und sei $q \approx 1$ die Häufigkeit des normalen Allels. Da p^2 sehr gering ist, kann p^2 für die folgende Berechnung vernachlässigt werden. Die Frequenz der heterozygoten Personen (d.h., mit einem normalen und einem mutierten Allel) beträgt $2pq \approx 2p$. Haben betroffene Personen einen Reproduktionsnachteil von s, dann ist der Anteil von Nachkommen von `Aa`-Individuen innerhalb der Population $2ps$ (der Anteil von $2p$ ist reduziert um den Faktor s). Da nur jeder zweite Nachkomme eines `Aa`-Individuums das `A`-Allel ererbt, reduziert sich der Anteil des Allels `A` um $\frac{1}{2} \times 2ps$. Ein Gleichgewicht besteht, falls diese Abnahme durch eine Neumutationsrate μ kompensiert wird, d.h.

$$\mu = ps \qquad \Rightarrow \qquad \hat{p} = \frac{\mu}{s}$$

Besteht ein vollständiger Reproduktionsnachteil, d.h., $s = 1$, dann gilt $\hat{p} = \mu$, die Häufigkeit des mutierten Allels in der Population ist gleich die Neumutationsrate. Besteht ein mäßiger Nachteil, z.B., $s = 1/3$, dann gilt $\hat{p} = 3\mu$.

3.2.1 Autosomal rezessive Erkrankungen

Betrachten wir nun eine autosomal rezessive Erkrankung, bei der homozygote Mutationsträger schwer erkrankt sind und nicht reproduzieren (Tabelle 3.4).

	`AA`	`Aa`	`aa`	Allelfrequenz von `a`
Vor Selektion	p^2	$2pq$	q^2	q
Nach Selektion	$\frac{p^2}{p^2+2pq}$	$\frac{2pq}{p^2+2pq}$	0	$\frac{q}{1+q}$

Tabelle 3.4: Selektion bei einer autosomal rezessiven lethalen Erkrankung.

Um die Veränderung in der Allelfrequenz q pro Generation zu berechnen, vergleichen wir die Häufigkeit von q vor und nach Selektion. `AA`- und `aa`-Individuen tragen keine Allele zur nächsten Generation bei und `Aa`-Individuen vererben ein `a`-Allel an durchschnittlich jeden zweiten Nachkommen. Daher ist die Häufigkeit vom `a`-Allel in der nächsten Generation $1/2 \times p(\texttt{Aa})$.

$$\begin{aligned} q_{n+1} &= \frac{1}{2} \times p(\texttt{Aa}) \\ &= \frac{1}{2} \times \frac{2p_n q_n}{p_n^2 + 2p_n q_n} \\ &= \frac{q_n}{p_n + 2q_n} \\ &= \frac{q_n}{1+q_n} \qquad \bullet \text{ da} \quad p+q=1 \end{aligned}$$

Wir können nun fragen, welche Auswirkung die Selektion bei autosomal rezessiven Erkrankungen auf die Häufigkeit von mutierten Allelen in der Population hat. Nehmen wir als Beispiel die seltene autosomal rezessive Erkrankung Galaktosämie (Prävalenz in der Bevölkerung 1:40.000 Menschen). Da $q^2 = 1/40.000$ folgern wir dass $q = 0,005$. Wie viele Generationen vergehen, bis sich die Häufigkeit q halbiert hat? Eine schnelle Berechnung in matlab zeigt:

```
f=@(q) q/(1+q);
q=0.005;
i=0;
while q>0.0025
  q=f(q);
  i=i+1;
end
display([ num2str(i) ' Generationen'])
```

$i = 200$ Generationen, oder ca. 6.000 Tausend Jahre unter der Annahme einer Generationszeit von 30 Jahren. Dies bedeutet, dass ein Versuch, die Häufigkeit von rezessiven Genmutationen in der Population zu verringern, indem homozygote Mutationsträger nicht reprozieren, extrem ineffizient wäre.

Aufgabe: Verwenden Sie das Hardy-Weinberg-Gesetz, um die Häufigkeit von gesunden heterozygoten Mutations(über)trägern bei der Galaktosämie zu bestimmen.

3.2.2 Sichelzellanämie

Ein Grund, weshalb die Mutationshäufigkeit in einer Population nicht abnimmt ist ein Heterozygotenvorteil. Ein gut untersuchtes Beispiel ist die Sichelzellanämie. Bei dieser autosomal rezessiven erblichen Erkrankung der roten Blutkörperchen (Erythrozyten) kommt es auf Grund einer Mutation im Gen für β-Globin zur Bildung eines abnormen Hämoglobins, wodurch sich die roten Blutzellen bei Sauerstoffarmut zu sichelförmigen Gebilden verformen, so dass diese durch Hämolyse abgebaut werden. Die klinische Erkrankung manifestiert sich nur bei homozygoten Mutationsträgern. Heterozygote Mutationsträger sind gegen Malaria resistent, weshalb die β-Globin-Mutation in Malariagebieten sehr verbreitet ist (Fig. 3.2).

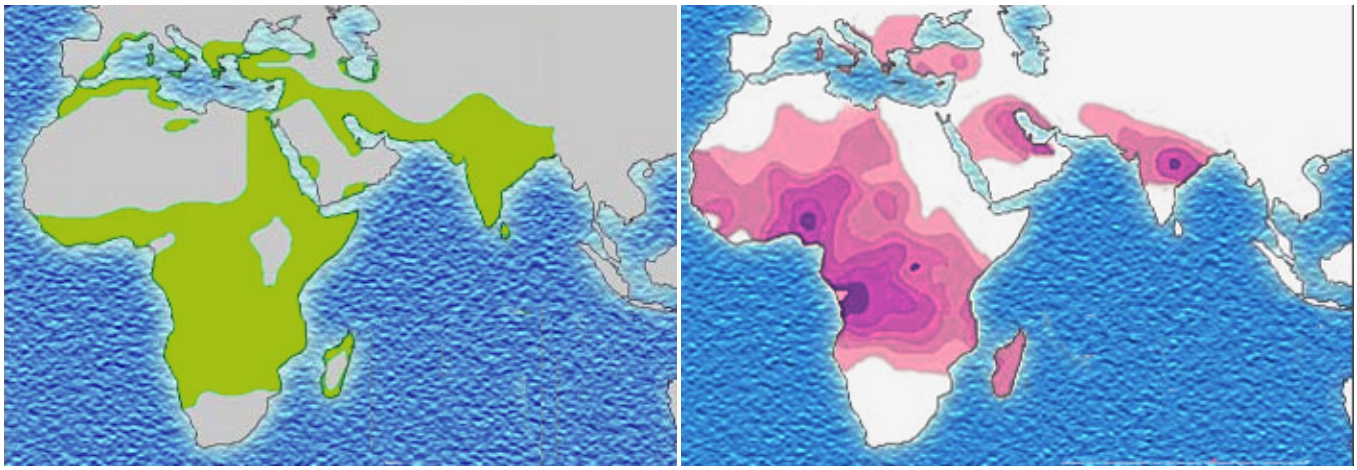


Abbildung 3.2: Links: Verbreitung der Malaria falciparum, rechts Verbreitung des Sichelzellallels. Bildquelle: Wikipedia Commons.

Um diese Situation zu modellieren, sei s der Selektionsnachteil von Personen ohne das Sichelzellallel (`AA`) und t der Selektionsnachteil von homozygoten Mutationsträgern (`aa`).

	`AA`	`Aa`	`aa`	Gesamt
Vor Selektion	p^2	$2pq$	q^2	1
Fitness	$1-s$	1	$1-t$	
Nach Selektion	$p^2(1-s)$	$2pq$	$q^2(1-t)$	$1-sp^2-tq^2$

Tabelle 3.5: Selektion zu Gunsten von Heterozygoten.

Gleichung 3.1 folgend[2] haben wir

$$\begin{aligned} q_{n+1} &= \frac{1/2 \times 2pq + q^2(1-t)}{(1-sp^2-tq^2)} \\ &= \frac{pq + q^2(1-t)}{1-sp^2-tq^2} \\ &= \frac{pq + q^2 - tq^2}{1-sp^2-tq^2} \\ &= \frac{q[p+q] - tq^2}{1-sp^2-tq^2} \\ &= \frac{q - tq^2}{1-sp^2-tq^2} \end{aligned}$$

Wir können die Veränderung von einer Generation zur nächsten berechnen als

$$\begin{aligned} \delta q_n &= q_{n+1} - q_n \\ &= \frac{q_n - tq_n^2}{1-sp_n^2-tq_n^2} - q_n \\ &= \frac{q_n - tq_n^2 - q_n + sp_n^2 q_n + tq_n^3}{1-sp_n^2-tq_n^2} \\ &= \frac{-tq_n^2 + sp_n^2 q_n + tq_n^3}{1-sp_n^2-tq_n^2} \\ &= \frac{q_n\left[tq_n\{q_n-1\} + sp_n^2\right]}{1-sp_n^2-tq_n^2} \\ &= \frac{q_n\left[tq_n\{-p_n\} + sp_n^2\right]}{1-sp_n^2-tq_n^2} \\ &= \frac{p_n q_n[sp_n - tq_n]}{1-sp_n^2-tq_n^2} \end{aligned}$$

Wir können fragen, ob q_n steigt oder sinkt von Generation zu Generation, d.h., welches Vorzeichen δq_n hat. Bei der Sichelzellanämie in einem bestimmten Malariagebiet hat man folgende Selektionsnachteile geschätzt:

- s (Selektion gegen Personen ohne das Sichelzellallel `AA`): 0,2039
- t (Selektion gegen homozygoten Mutationsträgern, `aa`): 0,8302

Wir plotten δq_n gegen q für alle möglichen Werte von q.

```
s=0.2039;
t=0.8302;

f=@(p,q,s,t) p*q*(s*p-t*q)/(1-s*p^2-t*q^2);

deltaq = [ ];

for q=0:0.001:1
    p=1-q;
    deltaq = [deltaq ; f(p,q,s,t) ];
end
clf
plot(0:0.001:1,deltaq);
line([0 1],[0 0]);
xlabel('q','FontSize',14);
ylabel('\Delta q','FontSize',14);
grid;
```

[2] `AA`-Individuen tragen keine `a`-Allele zur nächsten Generation bei. `Aa`-Individuen vererben im Durchschnitt das `a`-Allel an jeden zweiten Nachkommen, d.h., $2pq \times 1/2$ und `aa`-Individuen vererben ein `a`-Allel an jeden Nachkommen ($q^2(1-t) \times 1$).

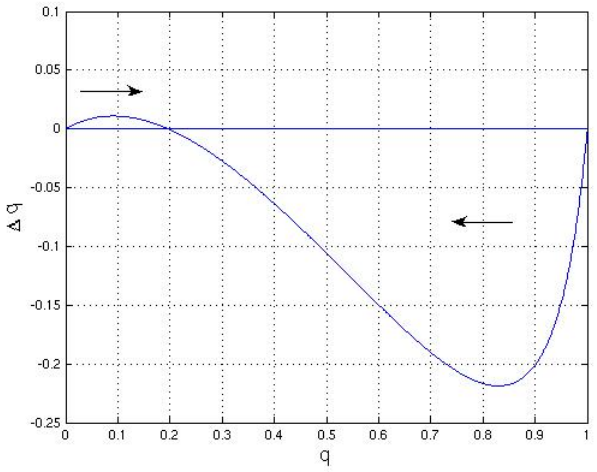


Abbildung 3.3: Veränderung der Allelfrequenz δq für unterschiedliche Werte von q bei der Sichelzellanämie. Für $q < 0,2$ steigt q von Generation zu Generation. Für $q > 0,2$ reduziert sich dieser Wert. Eine Heterozygotenfrequenz von bis 31,7% wurde in einigen Populationen in Malariagebieten gefunden, was in etwa $2pq = 2 \times 0,8 \times 0,2$ entspricht.

Kapitel 4

Kopplungsungleichgewicht

4.1 Rekombination

Definition 25 (Genotyp) Die (diploide) genetische Ausstattung eines Individuums an einem oder mehreren Loci. Beide Exemplare eines Allels werden berücksichtigt, z.B. der Genotyp an einem Locus mit Allelen `A` und `a` kann `AA`, `Aa` oder `aa` sein. □

Definition 26 (Haplotyp) Ein Block von Allelen, die mehr oder weniger eng benachbart auf demselben Chromosomenabschnitt liegen und eher zusammen (als durch Rekombination getrennt) als Block übertragen werden. □

Unter Rekombination versteht man im Allgemeinen den Austausch von Allelen und im Allgemeinen die Verteilung und Neuanordnung von genetischem Material. Die *interchromosomale Rekombination* bedeutet die Neukombination ganzer Chromosomen, was eine einfache Folge der unabhängigen Verteilung der einzelnen Chromosomen während der Meiose darstellt. Die *intrachromosomale Rekombination* entsteht durch Neukombination von Allelen innerhalb von Chromosomen durch *Crossing-over* in der Meiose I (Abb. 4.1). Im folgenden Text meinen wir mit "Rekombination" solange nichts anderes angegeben ist immer die intrachromosomale Rekombination

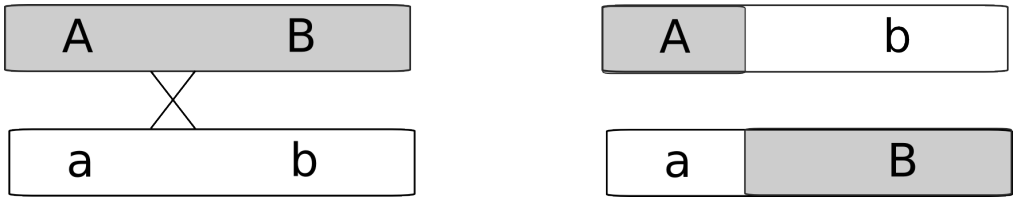


Abbildung 4.1: Intrachromosomale Rekombination. Links: Die Haplotypen sind `A-B` und `a-b`. Der Genotyp ist `Aa,Bb`. Die Allele sind `A,B,a` und `b`. Rechts: Nach einer Rekombination sind die Haplotypen `A-b` und `a-B`. Der Genotyp ist `Aa,Bb`. Die Allele sind `A,B,a` und `b`.

4.2 Kopplungsungleichgewicht

Ein Kopplungsungleichgewicht (häufig wird die engliche Abkürzung **LD** für *linkage disequilibrium* verwendet) besteht zwischen zwei Genloci, die auf einem Chromosom eng beieinander liegen, und deshalb zusammenvererbt werden. Ein Block von Loci, die zusammen vererbt werden, nennt man **Haplotyp**. Zwei eng beieinander liegende Loci werden dann nicht zusammen

vererbt, wenn zwischen ihnen eine Rekombination erfolgt. Die Analyse dieser Verhältnisse ist die Grundlage der Kopplungsanalyse. Im folgenden wollen wir jedoch erstmal ein allgemeines Modell für das LD beschreiben.

Sei `A` und `a` zwei Allele von einem Locus und `B` und `b` zwei Allele eines anderen Locus. Dann gibt es vier mögliche Kombinationen in den Gameten[1], und zwar `AB`, `Ab`,`aB` und `ab`. Seien die Häufigkeiten dieser Kombinationen innerhalb einer Population von Gameten p_{AB}, p_{Ab}, p_{aB} und p_{ab}. Dann sind die Allelfrequenzen:

$$p_a = p_{ab} + p_{aB} \qquad p_A = p_{Ab} + p_{AB} = 1 - p_a$$

und

$$p_b = p_{ab} + p_{Ab} \qquad p_B = p_{AB} + p_{aB} = 1 - p_b$$

Nimmt man an, dass die beiden Genloci untereinander im Kopplungsgleichgewicht sind (zum Beispiel weil die Genloci auf unterschiedlichen Chromosomen gelegen sind), dann ist die Wahrscheinlichkeit, dass eine Gamete das Allel `a` trägt unabhängig von der Wahrscheinlichkeit dass sie das Allel `b` trägt, d.h., es gilt $p_{ab} = p_a \times p_b$.

Sind die Loci nicht im Kopplungsgleichgewicht, dann gilt $p_{ab} \neq p_a \times p_b$. Wir führen die Variable D ein, um die Abweichung vom Kopplungsgleichgewicht zu beschreiben:

$$p_{ab} = p_a p_b + D \tag{4.1}$$

Hieraus folgt

$$p_{aB} = p_a - p_{ab} = p_a - p_a p_b - D = p_a(1 - p_b) - D = p_a p_B - D$$

Eine analoge Berechnung zeigt:

$$p_{Ab} = p_A p_b - D$$

$p_{ab} = p_a p_b + D$ ist eine Definition, und $p_{AB} = p_A p_B + D$ lässt sich wie folgt bestimmen

$$\begin{aligned} p_{AB} &= 1 - p_{ab} - p_{Ab} - p_{aB} = 1 - (p_a p_b + D) - (p_a p_B - D) - (p_A p_b - D) \\ &= 1 + D - (p_a p_b + p_a p_B + p_A p_b) = 1 + D - (1 - p_A p_B) \\ &= p_A p_B + D \end{aligned}$$

Die am häufigsten verwendete Definition der LD-Koeffiziente D ist jedoch

$$\boxed{D = p_{AB} p_{ab} - p_{Ab} p_{aB}} \tag{4.2}$$

Diese Formel leitet sich von den vorherigen ab:

$$\begin{aligned} D &= p_{ab} - p_a p_b \\ &= p_{ab} - (p_{aB} + p_{ab})(p_{Ab} + p_{ab}) \\ &= p_{ab} - p_{aB} p_{Ab} - p_{ab} p_{ab} - p_{aB} p_{ab} - p_{Ab} p_{ab} \\ &= p_{ab}(1 - p_{ab} - p_{aB} - p_{Ab}) - p_{aB} p_{Ab} \\ &= p_{ab}(p_{AB}) - p_{aB} p_{Ab} \end{aligned}$$

Die Koeffiziente D ist die einfachste von einer Reihe gebräuchlicher Maße für das LD. D kann schwer zu interpretieren sein. Das Vorzeichen ist willkürlich (es hängt davon ab, ob wir `A` und `B` für die häufigere Allele verwenden und `a` und `b` für die selteneren oder umgekehrt). Die Spannweite von D hängt von den Allelfrequenzen ab, so dass es schwierig ist, den Grad an LD zwischen zwei Markern miteinander zu vergleichen. Die Parameter D' und r^2 sind für bestimmte Anwendung vorteilhaft, sollen an dieser Stelle nicht weiter besprochen werden.

Abbildung 4.2 zeigt den Einfluss von unterschiedlichen Allelfrequenzen auf die Spannweite von D.

4.2.1 Woher kommt ein LD?

[1] zur Erinnerung sind Gameten Keimzellen, d.h. haploide Zellen, die im Gegensatz zu diploiden Zellen jeweils nur ein Exemplar jedes Locus haben.

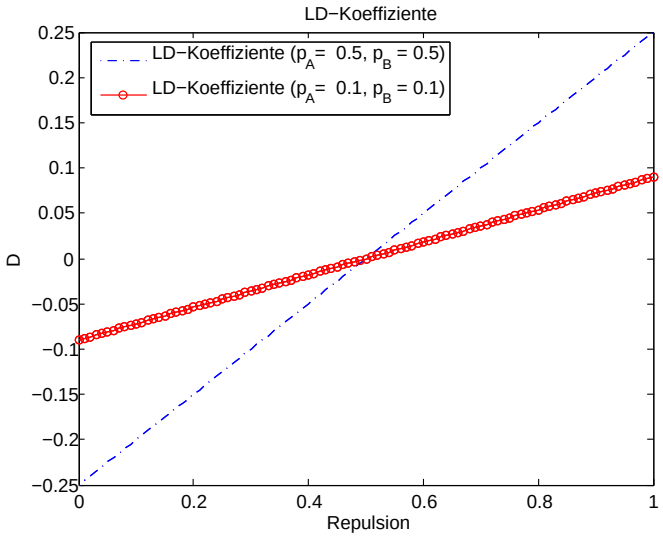


Abbildung 4.2: Abhängigkeit der LD-Koeffiziente D von den Allelfrequenzen und vom Grad an Repulsion. 0 = komplette Coupling von `AB` und `ab`, 1,0 = komplette Repulsion (Überschuss an `Ab` und `aB`).

Definition 27 (Gründereffekt) Gründereffekt (Founder-Effect) beschreibt eine genetische Abweichung einer Gründerpopulation von der Stammpopulation. □

Diese Abweichung entsteht aufgrund der geringen Anzahl an vorhandenen Allelen der an ihrer Gründung beteiligten Individuen und nicht infolge unterschiedlicher Selektionsbedingungen (Abb. 4.3). Zum Beispiel kann man Hinweise auf LD in den Bevölkerungen von Osterinsel oder der französischsprachigen Population von Quebec finden.

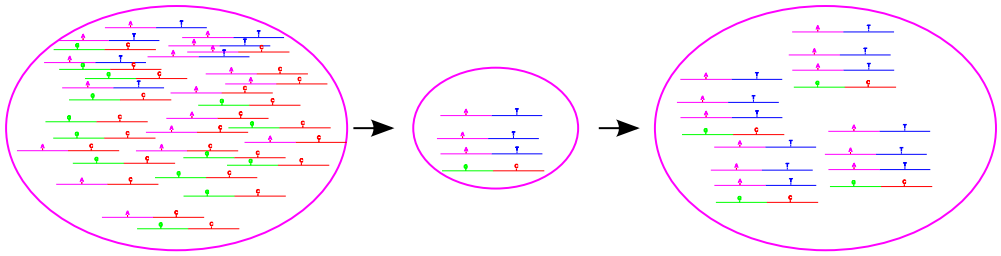


Abbildung 4.3: Gründereffekt und LD. Dargestellt sind jeweils zwei Genloci pro Individuum. Eine in der Stammpopulation bestehende, große genetische Variabilität (hier symbolisiert duch drei Haplotypen für die zwei Genloci, d.h. `G−C`, `A−C` unf `A−T` reduziert sich dadurch, dass eine neue Kolonie durch vier Individuen gegründet wird, bei denen nur zwei der drei Haplotypen vorhanden sind (`G−C` und `A−T`). Unter den Nachkommen dieser vier Individuen herrscht zunächst ein Kopplungsungleichgewicht zwischen beiden Genloci, da da Allel `A` an dem ersten Genort zunächst nur mit dem Allel `T` an dem zweiten Genort vorkommt. In den nachfolgenden Generation wird dieses Kopplungsungleichgewicht durch die Rekombination nach und nach verschwinden.

4.2.2 Wie lange dauert es, bis ein Kopplungsungleichgewicht verschwunden ist?

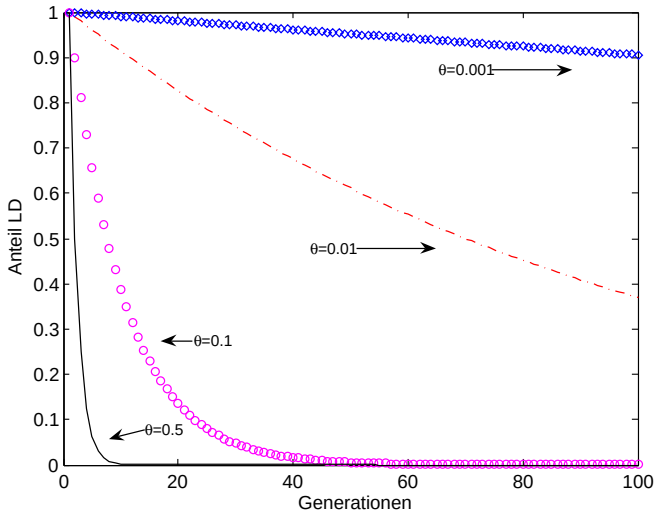
Sind die Loki nicht gekoppelt (z.B. weil sie auf zwei unterschiedlichen Chromosomen gelegen sind), dann gilt $\theta = 0,5$ und $D^{i+1} = \frac{1}{2}D^i$, d.h., das Kopplungsungleichgewicht halbiert sich mit jeder Generation.

Definition 28 (Rekombinationsfraktion) Die **Rekombinationsfraktion** zwischen zwei Loci ist der Anteil an postmeiotischen Chromosomen, bei denen eine Rekombination zwischen den Loci stattgefunden hat. Die Rekombinationsfraktion wird oft mit θ angegeben. Sie kann maximal 0,5 betragen. Ein $\theta < 0,5$ gilt als Hinweis für eine Kopplung zwischen zwei Loci. □

Es kann gezeigt werden, dass D mit jeder Generation um $(1-\theta)$ abnimmt. Dies impliziert:

$$D^i = (1-\theta)^i D^0$$

Die Entwicklung eines zunächst kompletten Kopplungsungleichgewichts wird für verschiedene Werte von θ in Abb. 4.2.2 gezeigt.



Normalisierende Selektion und Kopplungsungleichgewicht

Die natürliche Auslese wirkt häufig gegen Individuen an den Extremen des phänotypischen Spektrums und begünstigt Ausprägungen eines phänotypisches Merkmals, die dem Mittelwert des Merkmals in der Bevölkerung nahe sind. Dieses Phänomen ist zunächst Hermon Bumpus 1898 aufgefallen[2], nachdem er 136 Hausschwalben untersuchte, die nach einem ungewöhnlich schweren Schneesturm in das anatomische Labor der Brown University in Providence, Rhode Island, gebracht worden sind. 72 der Schwalben starben, 64 überlebten. Bumpus stellte fest, dass unter den überlebenden ein Überschuss an Vögeln mit durchschnittlichen Maßen hinlicklich Flügellänge vertreten waren, während Vögel mit kurzen oder langen Flügeln öfter als erwartet gestorben waren. Dies nennt sich **normalisierende Selektion**.

In der nachfolgenden Simulation wollen wir den Effekt der normalisierenden Selektion auf ein Paar von Genloci, welche beide additiv die Flügellänge bestimmen, untersuchen. Der erste Locus hat die Allele `A` und `a`, der zweite Locus `B` und `b`. Der Phänotyp bestimmt sich aus der Kombination von Allelen im Genotyp. Für jedes `a`- oder `b`-Allel ergibt sich 2 cm und für jedes `A`- oder `B`-Allel 3 cm. Zum Beispiel haben Vögel mit dem Genotyp `aabb` eine Flügellänge von $4 \times 2 = 8$ cm, solche mit dem Genotyp `aabB` $3 \times 2 + 3 = 9$ cm und solche mit `AABB` eine Flügellänge von $4 \times 3 = 12$ cm (Die Vögel mit einer Flügellänge von 10 cm haben eine höhere Fitness als diejenigen mit einer Flügellänge von 8 cm oder 12 cm (Tabelle 4.1).

Nehmen wir eine unendliche Population in Kopplungsgleichgewicht an, wobei die anfänglichen Allelfrequenzen $p(a) = 0,55$, $p(A) = 0,45$, $p(b) = 0,6$ und $p(B) = 0,4$ betragen. Dann gilt für die Gametfrequenzen $p(ab) = p(a)p(b)$ usw. Die Genotypfrequenzen können dann unter der Annahme der Zufallspaarung wie im folgenden Beispiel berechnet werden:

$$p(\frac{ab}{ab}) = p(ab)p(ab) \qquad p(\frac{aB}{ab}) = 2p(aB)p(ab)$$

[2]Bumpus, Hermon C. 1898. Eleventh lecture. The elimination of the unfit as illustrated by the introduced sparrow, Passer domesticus. (A fourth contribution to the study of variation.) Biol. Lectures: Woods Hole Marine Biological Laboratory, 209-225.

Phänotyp	8	9	10	11	12
Genotypen	$\frac{ab}{ab}$	$\frac{aB}{ab}$ $\frac{Ab}{ab}$	$\frac{aB}{aB}$ $\frac{aB}{Ab}$ $\frac{Ab}{Ab}$ $\frac{AB}{ab}$	$\frac{aB}{AB}$ $\frac{AB}{Ab}$	$\frac{AB}{AB}$
Fitness	0,8	0,9	1,0	0,9	0,8

Tabelle 4.1: Normalisierende Selektion.

Der Faktor 2 kommt daher, dass der Gamet aB von der Mutter und der Gamet ab vom Vater kommen kann oder auch umgekehrt.

Das Ergebnis der Simulierung, bei der eine Rekombinationsfrequenz von $\theta = 0,1$ zwischen den Genorten angenommen wurde, sind in Abb. 4.4 dargestellt.

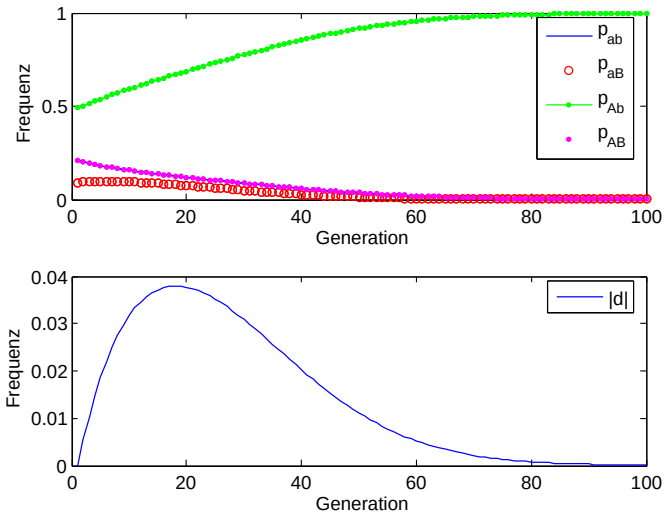


Abbildung 4.4: Normalisierende Selektion

Im folgenden wird der matlab-Code[3] erklärt, womit die Simulation durchgeführt wurde. Der erste Code-Abschnitt definiert die Häufigkeiten der Allele und der Haplotypen. Unter der Annahme eines Kopplungsgleichgewichts gilt $p(ab) = p(a)p(b)$ usw.

```
p_a=0.3;
p_b=0.7;
ngenerations=100;

p_A=1-p_a;
p_B=1-p_b;

%Initial besteht ein Kopplungsgleichgewicht, daher p(ab)=p(a)p(b) usw.
p_ab=p_a*p_b;
p_aB=p_a*p_B;
p_Ab=p_A*p_b;
p_AB=p_A*p_B;
```

Der nächste Code-Abschnitt definiert die Vektoren `d` als 100×1-Vektor und `genotype_freq` als 100×4-Matrix. Diese Variablen werden für Generationen $1 \ldots 100$ die Werte für die LD-Koeffiziente D und die Frequenzen der vier Genotypen festhalten. Die matlab-Funktion `zeros(M,N)` alloziert Speicher für eine $M \times N$-Matrix.

```
d= zeros(ngenerations,1);
genotype_freq = zeros(ngenerations,4);
```

[3] normalizingSelection.m, kann von http://compbio.charite.de heruntergeladen werden.

Im folgenden berechnen wir für die erste Generation D nach Gleichung 4.2 und speichern das Ergebnis im ersten Feld von d. Wir speichern die Genotypfrequenzen der ersten Generation in der ersten Reihe von genotype_freq.

```
D=p_AB*p_ab - p_Ab*p_aB;
d(1)=D;
genotype_freq(1,:)=[p_ab,p_aB,p_Ab,p_AB];
```

Wir nehmen an, dass ab diesem Punkt eine normalisierende Selektion herrscht. Der folgende Code berechnet den Effekt dieser Selektion unter der Annahme einer Rekombinationsfrequenz von $\theta = 0,1$ für 100 Generationen. Hierbei wird in der Funktion gtype_select ausgehend von den Genotypfrequenzen in Generation i die Frequenzen für die kommende Generation $i+1$ berechnet. Das Ergebnis wird in genotype_freq gespeichert und für die Berechnung von D für die aktuelle Generation verwendet (die Funktion gtype_select wird weiter unten erklärt).

```
for i=2:ngenerations
    % Calculate and store genotype frequencies
    [p_ab,p_aB,p_Ab,p_AB] = gtype_select(p_ab,p_aB,p_Ab,p_AB);
    genotype_freq(i,:)=[p_ab,p_aB,p_Ab,p_AB];
    %Calculate and store LD
    d(i)=p_AB*p_ab - p_Ab*p_aB;
end
```

Im nachfolgenden Code (hier nicht gezeigt) wird das Ergebnis geplottet (wie in Abb. 4.4).

Die Funktion gtype_select[4] definiert zunächst die Rekombinationsfrequenz θ und die in Tabelle 4.1 aufgeführten Werte für die relative Fitness.

```
function [p_ab,p_aB,p_Ab,p_AB] = gtype_select(p_ab,p_aB,p_Ab,p_AB)

%Rekombinationsfrequenz 0.1
theta=0.1;

%% Selektion auf Grund des Phaenotyps
%% 8-9-10-11-12 cm Fluegellaenge
fitness_8 = 0.8;
fitness_9 = 0.9;
fitness_10 = 1.0;
fitness_11 = 0.9;
fitness_12 = 0.8;
```

Dann werden aus den Haplotypfrequenzen die (diploiden) Genotypfrequenzen berechnet. Diese Frequenzen werden durch die relative Fitness multipliziert. Zum Beispiel ist die Häufigkeit des Genotyps `ab/ab` $p(ab)p(ab)$, die Fitness ist 0,8 (dieser Genotyp hat eine geringere Fitness und trägt weniger Gameten zur nächsten Generation bei). Der Faktor 2 in der Berechnung der Genotypfrequenz p_ab_aB kommt von der Tatsache, dass der Haplotyp `ab` von der Mutter oder vom Vater kommen kann (wie beim Hardy-Weinberg-Gesetz).

```
%% phenotype = 8 cm
p_ab_ab=p_ab^2 * fitness_8;

%% phenotype = 9 cm
p_ab_aB = 2*p_ab*p_aB * fitness_9;
p_ab_Ab = 2*p_ab*p_Ab * fitness_9;

%% phenotype = 10 cm
p_Ab_aB = 2* p_Ab * p_aB * fitness_10;
p_AB_ab = 2*p_AB*p_ab * fitness_10;
p_Ab_Ab = p_Ab^2 * fitness_10;
p_aB_aB = p_aB^2 * fitness_10;

%% phenotype = 11 cm
p_Ab_AB = 2*p_Ab*p_AB * fitness_11;
p_aB_AB = 2*p_aB*p_AB * fitness_11;

%% phenotype = 12 cm
p_AB_AB = p_AB^2 * fitness_12;
```

Die Summe der einzelnen Häufigkeiten muss 1 ergeben, weshalb wir renormalisieren müssen:

```
%Renormalize
total = p_ab_ab + p_ab_aB + p_ab_Ab + p_Ab_aB + p_AB_ab + p_Ab_Ab ...
    + p_aB_aB + p_Ab_AB + p_aB_AB + p_AB_AB;
```

[4] auch von http://compbio.charite.de herunterzuladen.

```
p_ab_ab = p_ab_ab / total;
p_ab_aB = p_ab_aB / total;
p_ab_Ab = p_ab_Ab / total;
p_Ab_aB = p_Ab_aB / total;
p_AB_ab = p_AB_ab / total;
p_Ab_Ab = p_Ab_Ab / total;
p_aB_aB = p_aB_aB / total;
p_Ab_AB = p_Ab_AB / total;
p_aB_AB = p_aB_AB / total;
p_AB_AB = p_AB_AB / total;
```

Nun können wir die Genotypfrequenzen der Gameten berechnen, woraus die nächste Generation entsteht. Einige, aber nicht alle Rekombinationen führen zu neuen Haplotypen[5]:

$$\begin{aligned}
&\text{Genotyp} \rightarrowtail \text{Gameten} \\
&\boxed{AB/AB} \overset{\theta}{\rightarrowtail} \boxed{AB} \,\&\, \boxed{AB} \\
&\boxed{ab/Ab} \overset{\theta}{\rightarrowtail} \boxed{Ab} \,\&\, \boxed{ab} \\
&\boxed{aB/ab} \overset{\theta}{\rightarrowtail} \boxed{ab} \,\&\, \boxed{aB} \\
&\boxed{aB/Ab} \overset{\theta}{\rightarrowtail} \boxed{ab} \,\&\, \boxed{AB} \\
&\boxed{ab/ab} \overset{\theta}{\rightarrowtail} \boxed{ab} \,\&\, \boxed{ab}
\end{aligned}$$

Wir können nun die Frequenz des Haplotyps `ab` unter den Gameten berechnen als

$$\begin{aligned}
p(ab) &= p\left(\frac{ab}{ab}\right) && \bullet \text{ Jeder Gamet, der von einem } \texttt{ab/ab}\text{-Individuum kommt, hat den Genotyp } \texttt{ab} \\
&+ 0.5 \times p\left(\frac{ab}{aB}\right) && \bullet \text{ Erst jeder zweite Gamet hat den Genotyp } \texttt{ab} \\
&+ 0.5 \times p\left(\frac{ab}{Ab}\right) && \bullet \text{ Erst jeder zweite Gamet hat den Genotyp } \texttt{ab} \\
&+ (1-\theta) \times 0.5 \times p\left(\frac{AB}{ab}\right) && \bullet \text{ Jeder 2. nicht rekombinante Gamet hat den Genotyp } \texttt{ab} \\
&+ \theta \times 0.5 \times p\left(\frac{Ab}{aB}\right) && \bullet \text{ Jeder 2. rekombinante Gamet hat den Genotyp } \texttt{ab}
\end{aligned}$$

Die Berechnungen für die übrigen drei Gametengenotypen erfolgen analog.

```
p_ab = p_ab_ab ...
     + 0.5 * p_ab_aB ...
   + 0.5 * p_ab_Ab ...
   + (1-theta) * 0.5 * p_AB_ab ...
   + theta * 0.5 * p_Ab_aB;

p_aB = 0.5 * p_ab_aB ...
    + 0.5 * (1-theta)* p_Ab_aB ...
    + p_aB_aB ...
    + 0.5*p_aB_AB...
      + 0.5*theta*p_AB_ab;

p_Ab = 0.5 *p_ab_Ab ...
    + 0.5 * (1-theta)*p_Ab_aB ...
    + p_Ab_Ab ...
    + 0.5 * p_Ab_AB ...
    + 0.5*theta*p_AB_ab;

```

[5] Bemerke, dass wir der Einfachheit halber die Rekombination so modellieren, dass bei einer Rekombination alle Chromatiden rekombinieren und nicht nur zwei der vier Chromatiden (vgl. Abb. 2.11 von Strachan und Read)

```
19 p_AB = 0.5 *(1-theta)* p_AB_ab ...
20     + 0.5* p_Ab_AB ...
21     + 0.5*p_aB_AB ...
22     + p_AB_AB ...
23     + theta * 0.5 * p_Ab_aB;
```

Wie in Abb. 4.4 ersichtlich, kommt es dazu, dass fast jedes Individuum den Genotyp `Ab/Ab` hat und somit den günstigsten Phänotyp (Flügellänge 10 cm). Dies führt dazu, dass D auf Null sinkt, obwohl es eine komplette Kopplung der Allele `A` und `b` gibt.

Ein Beispiel der normalisierenden Selektion in menschlichen Gesellschaften ist die Beobachtung, dass Neugeborene mit einem besonders leichten oder schweren Geburtsgewicht (zumindest vor dem Zeitalter der modernen Neonatologie) eine erhöhte Sterblichkeit aufwiesen. Die normalisierende Selektion ist einer von mehreren Faktoren, wie auch Beimischung (admixture) zweier Bevölkerungsgruppen und genetische Drift, der zum Kopplungsungleichgewicht führen kann. In einer späteren Vorlesung werden wir auf die zentrale Bedeutung des Kopplungsungleichgewichts für *Assoziationsstudien*, die zur Kartierung von Genen, welche mit einer erhöhten Anfälligkeit für Krankheiten (wie z.B. Bluthochdruck) verbunden sind, eingehen.

Kapitel 5

Anhang

Dieser Anhang enthält ergänzendes oder erklärendes Material, das über die Lernziele der Vorlesungsreihe hinausgeht (und daher für die Prüfung unrelevant ist), jedoch als Vorschau auf künftige Themen bzw. für spätere Referenz hier aufgeführt wird.

5.1 Die logistische Gleichung

Wir beschreiben das Wachstum einer Population mit

$$\frac{\mathrm{d}x}{\mathrm{d}t} = rx(1-x/K)$$

wobei K die Kapazität beschreibt. Wir haben

$$\frac{\mathrm{d}x}{x(1-x/K)} = r\mathrm{d}t$$

Zerlegt man den Integranden auf der linken Seite in partielle Brüche, erhält man:

$$\frac{\mathrm{d}x}{x} + \frac{1/K\mathrm{d}x}{1-x/K} = r\mathrm{d}t$$

Da $\int \frac{\mathrm{d}x}{x} = \log x + C_1$ und $\int \frac{1/K\mathrm{d}x}{1-x/K} = -\log(1-x/K) + C_2$, kann man folgern dass:

$$\begin{aligned} \int \left[\frac{\mathrm{d}x}{x} + \frac{1/K\mathrm{d}x}{1-x/K}\right] &= \int r\mathrm{d}t \\ \log x + C_1 - \log(1-x/K) + C_2 &= rt + C_3 \\ \log\left(\frac{x}{1-x/K}\right) &= rt + k \end{aligned}$$

wo die willkürlichen Integrationskonstanten kombiniert wurden: $k = C_1 + C_2 + C_3$. Man exponentiiere beide Seiten:

$$\frac{x}{1-x/K} = e^{rt+k} = Ce^{rt} \tag{5.1}$$

wo $C = e^k$. Es folgt:

$$\begin{aligned} \frac{x}{1-x/K} &= Ce^{rt} \\ x &= (1-x/K)Ce^{rt} \\ x + \frac{x}{K}Ce^{rt} &= Ce^{rt} \\ x\left(1+Ce^{rt}/K\right) &= Ce^{rt} \\ x &= \frac{Ce^{rt}}{1+Ce^{rt}/K} \end{aligned}$$

Um den Wert von C zu berechnen, setzt man $t = 0$ in Gleichung 5.1:

$$\begin{aligned} \frac{x_0}{1 - x_0/K} &= Ce^0 \\ C &= \frac{Kx_0}{K - x_0} \end{aligned}$$

5.2 Ein ODE-Modell für die Konkurrenz zwischen zwei Bakterienspezies

Gewöhnliche Differentialgleichungen (englisch ordinary differential equation, ODE) stellen ein häufig verwendetes mathematisches Werkzeug dar, um das Änderungsverhalten von Größen über eine Zeitspanne zu untersuchen. Sei $y(t)$ der Wert (die Werte) der Varibalen zum Zeitpunkt t und f eine Funktion, welche $\frac{dy}{dt}$ berechnet, d.h., $y'(t) = f(t, y(t))$ Solche Gleichungen haben üblicherweise viele Lösungen, jedoch kann durch Angabe von Anfangsbedingungen zu Zeitpunkt t_0 eine eindeutige Lösung festgelegt werden. Häufig sind unterschiedliche Anfangsbedingen mit unterschiedlichen Trajektorien (”Bahnen”) im Lösungsraum assoziiert.

In matlab wird die Funktion $f(t,y)$ in einer m-Datei angegeben, wobei das Argument t einen SKalarwert (Zeit) und das Argument y einen $(n \times 1)$ Zustandsvektor darstellt (Werte des Systems zu Zeitpunkt t). Die Funktion hat den zugehörigen Vektor f(t,y)=$\frac{dy}{dt}$ zurückzuliefern. Die m-Datei für das im Abschnitt 2.3 beschriebene System lautet:

```
1  function ydot = growth(t,y)
2
3  r1=0.1;  %growth constant for x
4  r2=0.4;  %growth constant for y
5  K1=100;  %capacity for x;
6  K2=60;   %capacity for y
7
8  x=y(1)
9  y=y(2)
10
11 ydot = zeros(2,1);
12 dxdt = r1*x*(1-(x+y)/K1);
13 dydt = r2*y*(1-(x+y)/K2);
14 ydot(1)=dxdt;
15 ydot(2)=dydt;
```

Hier werden Werte für die Konstanten in Gleichungen (2.9) und (2.10) angegeben (Zeilen 3–6), Werte für x und y aus dem Argument y extrahiert (Zeilen 8–9), und Speicher für den Rückgabewert alloziert (Zeile 10). Zeilen 11 und 12 berechnen Gleichungen (2.9) und (2.10). Die Funktion muss einen Vektor ydot zurückgeben, welcher dieselbe Größe und Belegung hat wie der Eingabevektor y.

Um eine Lösung zu finden, geben wir einen Zeitraum (Anfang bis Ende) sowie Anfangswerte an:

```
t = [0 100];              % Start- bis Endzeitpunkt
y0 = [10 10];             % Initialwerte  x=y0(1) und y=y0(2)
```

Wir rufen den ODE-Löser auf (es gibt mehrere eingebaute matlab-Funktion zur Lösung von ODEs, ode45 ist ein guter allgemeiner Algorithmus für die meisten Probleme):

```
Y=ode45(@growth,t,y0);
```

Und plotten das Ergebnis (vgl. Abb.

```
clf;                      % clear any previous figure
plot(Y.x,Y.y(1,:),'r-');
hold on;
plot(Y.x,Y.y(2,:),'b--');
legend('x','y');
xlabel('Zeit');
ylabel('Zellzahl');
```

Bemerkung: dieser Code funktioniert nicht ohne weiteres in octave, da die Syntax des ODE-Lösers etwas unterschiedlich ist. Um den Code anzupassen, muss man die Signatur der Funktion `growth` ändern zu:

```
function ydot = growth(y,t)
```

(d.h., die Reihenfolge der Argumente wird vertauscht, der restliche Code bleibt unverändert). Um das ODE-System zu lösen, ist folgender Code notwendig:

```
t=linspace(0,100,100);
y0 = [10 10];
Y=lsode('growth',y0,t);
plot(Y);
```

5.3 Ein matlab-Skript, um die genetische Drift darzustellen

Im folgenden werden das matlab/octave-Skript `gdrift.m` erklärt. Studenten sollen insbesondere die folgenden drei Variablen verändern, um den Einfluss der Parameter Populationsgröße (`N`) und Zeit (`ngenerationen`) auf die genetische Drift zu untersuchen.

```
N=100; N=100; % Populationsgroesse
p=0.5; % Anteil Typ A in Population
ngenerationen = 200;
```

matlab/octave ist eine High-Level-Computersprache, d.h., die Abstraktionsebene liegt (im Gegensatz etwa zu C) weit oberhalb der Maschinenebene. Das folgende Kommando vereinigt mehrere Operationen.

```
pop=rand(N,1)>0.5;
```

Das Kommando rand(1,N) vereinbart einen $1 \times N$-Vektor und initialisiert die Elemente mit Zufallszahlen $\in (0,1)$. Das Kommando rand(1,N)>0.5 prüft dann die so initialisierten Element mit 0.5 und liefert das Ergebnis dieses Vergleiches als 0 (falsch) bzw. 1(wahr) zurück. Zum Beispiel

$$\boxed{rand(3,1)} = \begin{bmatrix} 0.44725 \\ 0.54643 \\ 0.44388 \end{bmatrix}$$

und

$$\boxed{rand(3,1) > 0.5} \Rightarrow \begin{bmatrix} 0.44725 > 0.5 \\ 0.54643 > 0.5 \\ 0.44388 > 0.5 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Das folgende Kommando initialisiert einen Vektor `A`. Der Vektor hat `ngenerations` Reihen und 1 Spalte. Das Element `A(i)` wird jeweils die Anzahl von Individuen mit einem A-Allel in Generation `i` enthalten (die Anzahl der Individuen mit einem a-Allel ist dann `N-A(i)`).

```
A=zeros(ngenerations,1);
```

Die folgende For-Schleife wird `ngenerations`-mal durchgeführt und berechnet jeweils die Anzahl von Individuen in der Bevölkerung mit einem A-Allel in Generation `i`.

```
for i=1:ngenerations
    A(i) = sum(pop);
    ind=ceil(N*rand(N,1));
    pop=pop(ind);
end
```

Der Vektor pop hat N Elemente (eins für jedes Individuum in der Bevölkerung). pop(i) ist 1, falls Individuum i ein A-Allel bzw. 0, falls Individuum i ein a-Allel hat. Das Kommando sum(pop) berechnet die Summe aller Elemente von pop, was in diesem Fall dasselbe ist wie die Anzahl der A-Allele ist.

Das nächste Kommando verbindet 3 Schritte in einer Zeile. rand(N,1) haben wir oben kennengelernt. N*rand(N,1) multipliziert die Zufallszahlen $\in (0,1)$ mal N, was Zahlen $\in (0,N)$ erzeugt. Das ceil-Kommando erhöht jede reelle Zahl auf die nächstgrößere Ganzzahl. Zum Beispiel

$$\boxed{rand(3,1)} = \begin{bmatrix} 0.039425 \\ 0.762976 \\ 0.432788 \end{bmatrix} \Rightarrow \boxed{3 * rand(3,1)} = \begin{bmatrix} 0.11827 \\ 2.28893 \\ 1.29837 \end{bmatrix} \Rightarrow \boxed{ceil(3 * rand(3,1))} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$

Da matlab (im Gegensatz zu den meisten Computersprachen wie z.B. C und Java) die Indizierung von Arrays mit 1 anfangen lässt, können wir diese Zahlen als zufällige Vektorindices verwenden (da sie zwischen 1 und N liegen). Nach dem Kommando ind=ceil(N*rand(N,1)); können wir also pop verwenden, um die Eltern der Individuen in der nächsten Generation per Zufall auszuwählen. `pop(j)` enthält den Index des Elternteils von Individuum `j`. Mit dem nächsten Kommando pop=pop(ind) vertauschen

wir die Indices. Zum Beispiel könnte pop in Generation i folgende Werte für eine Population mit 5 Individuen enthalten:

$$pop = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Das Ergebnis von ind=ceil (N∗rand(N,1)); könnte folgendermaßen ausfallen:

$$ind = \begin{bmatrix} 4 \\ 5 \\ 1 \\ 3 \\ 2 \end{bmatrix}$$

Das Kommando pop(ind) liefert die Element von pop nach der Reihenfolge von ind zurück, so

$$pop(ind) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Das Ergebnis wird schließlich wieder der Variablen `pop` für die nächste Generation zugewiesen Schließlich wird das Ergebnis geplottet:

```
clf; % clear current figure.
gen = [1:ngenerations]';
plot(gen,A,'ro-');

axis([ 0 ngenerations 0 100]); % Achsen einstellen [xmin xmax ymin ymax]
xlabel('Generation');       % Beschriftung der X-Achse
ylabel('Anteil der Individuen mit Allel A');
title('Genetische Drift');
```