

medgen 2008 · 20:282–287
 DOI 10.1007/s11825-008-0115-x
 Online publiziert: 1. September 2008
 © Springer Medizin Verlag 2008

A. Caliebe

Institut für Medizinische Informatik und Statistik, Haus 31,
 Universitätsklinikum Schleswig-Holstein, Campus Kiel, Kiel

Mathematische Modelle in der Populationsgenetik

Die klassische mathematische Populationsgenetik ist eine Synthese aus mendelscher Genetik und darwinscher Evolutionstheorie. Ihr Ursprung liegt in den grundlegenden Arbeiten von Fisher [4, 5], Wright [15] und Haldane [6, 7] aus den 1920er und 1930er Jahren. Durch die Einführung konkreter mathematischer Modelle für die zeitliche Entwicklung von Populationen unter Einbeziehung ihrer genetischen Grundlagen gelang es erstmals, nicht nur qualitative Aussagen über evolutionäre Zusammenhänge zu treffen, sondern hierzu auch quantitative Hypothesen zu formulieren, die sich anhand empirischer Daten überprüfen lassen. Die wesentlichen Mechanismen, die in diesen frühen mathematischen Modellen Berücksichtigung fanden, waren Selektion und Mutation. Dabei standen zunächst deterministische Modelle für die Entwicklung der Allelfrequenzen in unendlich großen Populationen im Vordergrund, in denen zufällige Einflüsse durch Reproduktionsschwankungen unberücksichtigt blieben. Auch Selektion und Mutation wurden als deterministische Prozesse behandelt. Diese Modelle lieferten trotz ihrer Einfachheit viele wichtige Resultate, mit denen sich z. B. das Entstehen und die Aufrechterhaltung von Polymorphismen in Populationen erklären ließen.

Im Wright-Fisher-Modell [5, 15] wurde erstmals der Effekt der Endlichkeit von Populationen berücksichtigt. Die durch die Endlichkeit bedingte stochastische Natur des Reproduktionsmechanismus hat einen wesentlichen Einfluss auf den Polymorphiegrad der Gene in der Population. Sie drückt sich mathematisch im Vorliegen eines Markov-Prozesses aus, d. h. der Polymorphiegrad einer Genera-

tion wird z. T. aus dem Polymorphiegrad der vorigen Generation, z. T. aber auch durch zufällige Einflüsse bestimmt. Neben der Fortpflanzung lassen sich auch Selektion und Mutation als zufällig modellieren, und bei der Analyse populationsgenetischer Daten ist mittlerweile die Verwendung stochastischer Modelle zum Standard geworden. Dies ist einerseits dadurch zu erklären, dass die Relevanz der Endlichkeit von Stichproben und Populationen, insbesondere wegen der Mög-

lichkeit so genannter „bottlenecks“, klar erkannt wurde. Andererseits sind Informationen über die zufälligen Schwankungen einer genetischen Populationsstruktur notwendig, um die Genauigkeit der Schätzung von Parametern und die Validität statistischer Tests beurteilen zu können.

Der vorliegende Artikel soll einen ersten Einblick in die grundlegenden stochastischen Modelle der Populationsgenetik vermitteln. Dabei wird der Schwer-



Abb. 1 ▲ Ronald A. Fisher. (Aus: <http://www-history.mcs.st-andrews.ac.uk/Mathematicians/Fisher.html>)



Abb. 2 ▲ Sewall Wright. (Aus: http://cdm.ampilsoc.org/cdm4/item_viewer.php?CISOROOT=%2Fgenetics&CISOPTR=125&DMSCALE=100&DMWIDTH=600&DMHEIGHT=600&DMMODE=viewer&DMFULL=1&DMX=0&DMY=0&DMTEX=&DMTHUMB=1&REC=8&DMROTATE=0&x=115&y=121)

punkt mehr auf ein qualitatives Verständnis der Modelle gelegt als auf Vollständigkeit oder quantitative Genauigkeit. Es versteht sich von selbst, dass daher viele der äußerst wichtigen weiterführenden Modelle, die allgemeinere oder speziellere Reproduktions-, Mutations- und Selektionsmodelle einschließen oder die Effekte ethnischer oder geografischer Strukturierung berücksichtigen, hier nicht behandelt werden können.

Wright-Fisher-Modell

Das klassische Modell für die Reproduktion einzelner Gene in einer endlichen Population ist das Wright-Fisher-Modell. Es wurde nahezu zeitgleich von Ronald A. Fisher (■ **Abb. 1**) und Sewall Wright (■ **Abb. 2**) entwickelt und untersucht [5, 15] und kann als Ausgangspunkt der modernen Populationsgenetik angesehen werden. Seine wesentlichen Annahmen sind eine zeitlich unveränderte Populationsgröße von N diploiden Individuen, eine nicht überlappende Generationenfolge und das so genannte „random mating“, d. h. eine im Hinblick auf das betrachtete Gen zufällige Wahl der Fortpflanzungspartner. Aus den $2N$ Genen einer Generation entsteht die Folgegeneration durch Ziehen mit Zurücklegen, d. h. jedes der $2N$ Gene der neu entstehenden Generation „wählt“ sich zufällig seinen Vorläufer unter den Genen der Elterngeneration aus. Dabei können manche Gene der Elterngeneration mehr als einen Nachkommen unter den Genen der nachfolgenden Generation haben.

Der hier skizzierte Reproduktionsprozess erzeugt eine Genealogie der $2N$ Gene in aufeinander folgenden Generationen. ■ **Abb. 3a** zeigt ein Beispiel einer solchen Genealogie. Man beachte, dass z. B. Gen 3 in Generation 4 drei Nachkommen in Generation 5 hat, wohingegen sich die Gene 4 und 5 der Generation 4 nicht fortgepflanzt haben. Das Wright-Fisher-Modell ist in biologischer Hinsicht höchst unrealistisch, in mathematischer Hinsicht jedoch sehr einfach. Obwohl es nur eine grobe Approximation der Realität darstellt, erlaubt es die Einbindung verschiedenster Selektions- und Mutationsmechanismen und vermittelt dadurch viele grundlegende Prinzipien der Popu-

lationsgenetik, insbesondere des Effekts endlicher Populationsgrößen.

2 Allele

Im folgenden Abschnitt wird das Wright-Fisher-Modell für ein Gen mit 2 Allelen, „A“ und „a“ betrachtet. Es wird zwischen 3 möglichen Szenarien der Evolution dieses Gens unterschieden, je nachdem ob Mutationen oder Selektion berücksichtigt werden.

Keine Mutation, keine Selektion

Ohne Mutation und Selektion kommen Generationsunterschiede in den Allelfrequenzen allein durch die Zufälligkeit der „Auswahl“ der Eltern der Gene zustande. In jeder Generation folgt die Anzahl der Gene, die das Allel A aufweisen, einer Binomialverteilung. Deren Erfolgswahrscheinlichkeit ist in der n . Generation durch die Frequenz π_{n-1} des Allels A in der Elterngeneration (d. h. der Generation $n-1$) bestimmt. Da keine Mutationen auftreten, muss langfristig eins der beiden Allele aussterben, während das andere fixiert wird. Ausgehend von der Anfangspopulation (d. h. der 0. Generation) beträgt die Wahrscheinlichkeit, mit der das Allel A fixiert wird, genau π_0 (von nun an kurz: π). Dies lässt sich dadurch erklären, dass nach hinreichend langer Zeit alle Gene in der Population von einem einzigen Gen der Ausgangspopulation abstammen. Die Wahrscheinlichkeit, dass dieses Gen das Allel A aufwies, entspricht der Frequenz π des Allels A in der Ausgangspopulation [3].

Neben der Wahrscheinlichkeit für die Fixierung von A interessieren sich Populationsgenetiker oft auch für die erwartete, bis dahin vergehende Zeit $E(\tau)$ (unter der Bedingung, dass A und nicht a fixiert wird). Hierfür fanden Ohta u. Kimura [14] die folgende Approximation:

$$E(\tau) \approx \frac{1}{\pi} 4N(1-\pi) \ln(1-\pi).$$

Man sieht an dieser Formel, dass die Zeit bis zur Fixierung eines Allels mit sinkender Anfangsfrequenz und wachsender Populationsgröße steigt. Tritt in einer Population ein neues, selektiv neutrales Allel eines Gens auf, wird dieses mit einer Wahrscheinlichkeit von $1/(2N)$

medgen 2008 · 20:282–287
DOI 10.1007/s11825-008-0115-x
© Springer Medizin Verlag 2008

A. Caliebe

Mathematische Modelle in der Populationsgenetik

Zusammenfassung

In der mathematischen Populationsgenetik wird der Einfluss von Selektion und Mutation auf die zeitliche Entwicklung der genetischen Struktur einer Population modelliert. Da alle Populationen und erst recht die für ihre Untersuchung zur Verfügung stehenden Stichproben endlich sind, spielen die stochastischen Aspekte der verwendeten Modelle eine besondere Rolle. Wichtige populationsgenetische Modelle sind das Wright-Fisher-Modell und das Koaleszenzmodell für die Genealogie von Stichproben sowie das „infinite alleles model“ und das „infinite sites model“ für die darin ablaufenden Mutationsprozesse.

Schlüsselwörter

Wright-Fisher-Modell · Koaleszenz · „infinite alleles model“ · „infinite sites model“ · Genetische Drift

Mathematical models in population genetics

Abstract

In mathematical population genetics, the influence of selection and mutation on the evolution of a population is modelled. Because all populations and particularly the samples used for their analysis are finite, the stochastic nature of these models plays an important role. Relevant genetic models include the Wright-Fisher model and the coalescence model for the genealogy of samples, as well as the infinite alleles model and the infinite sites model for the mutation processes superimposed upon these genealogies.

Keywords

Wright-Fisher model · Coalescence · Infinite alleles model · Infinite sites model · Genetic drift

Tab. 1 Fixierungswahrscheinlichkeit p des Allels A im Wright-Fisher-Modell bei gerichteter Selektion

Populationsgröße	Selektionskoeffizient	Ausgangsfrequenz von Allel A ^a	Fixierungswahrscheinlichkeit
N	s	π	p
10 ⁴	10 ⁻³	0,5	>0,9999
10 ⁴	10 ⁻³	0,05	0,6321
10 ⁴	10 ⁻³	1/2N=5×10 ⁻⁵	1,00×10 ⁻³
10 ³	10 ⁻³	0,5	0,7311
10 ³	10 ⁻³	0,05	0,1101
10 ³	10 ⁻³	1/2N=5×10 ⁻⁴	1,16×10 ⁻³
10 ³	10 ⁻⁴	0,5	0,5250
10 ³	10 ⁻⁴	0,05	0,0549
10 ³	10 ⁻⁴	1/2N=5×10 ⁻⁴	5,52×10 ⁻⁴

^aEntspricht gleichzeitig der Fixierungswahrscheinlichkeit für s=0

fixiert, wobei die zur Fixierung benötigte Zeit im Durchschnitt nach obiger Formel $2N \times 4N \times (1 - 1/2N) \times \ln(1 - 1/2N) \approx 4N$ Generationen beträgt.

Das Phänomen der zufälligen Schwankung von Allelfrequenzen in endlichen Populationen wird als *genetische Drift* bezeichnet. Sie spielt für die Evolution einer Population gegenüber der Selektion eine umso größere Rolle, je kleiner die Population ist. Hierbei können auch „bottlenecks“ und eine geringe Anzahl von Gründungsindividuen einer Population von Bedeutung sein. Nach der Neutralitätstheorie von Kimura [8] wird das Zustandekommen genetischer Variation sogar im Wesentlichen durch die genetische Drift selektiv neutraler Neumutationen in endlichen Populationen erklärt. Obwohl jede einzelne Mutation nur eine geringe Wahrscheinlichkeit hat, sich in höherer Frequenz in einer Population anzureichern, erweist sich bei gleichzeitiger Betrachtung aller Loci eine erhebliche Anzahl von diesen als polymorph durch die sehr hohe Anzahl an Mutationen an verschiedenen Loci. Man beachte jedoch, dass diese Polymorphismen keine stationären Zustände sind, sondern einen Übergang bis zur Fixierung eines der jeweiligen Allele darstellen. Die Zeit bis dahin ist allerdings mit $4N$ Generationen gerade in den heutigen Humanpopulationen sehr groß. Somit bilden die meisten der gegenwärtig existierenden Polymorphismen im Lichte der Neutralitätstheorie einen historischen Schnappschuss, der im Zuge der anschließenden raschen Expan-

sion der Erdbevölkerung quasi „eingefroren“ wurde.

Mutation

Im Folgenden nehmen wir an, dass das Allel A im Wright-Fisher-Modell bei der Weitergabe von einer Generation zur nächsten mit einer Rate v in das Allel a mutiert und dass das Allel a in gleicher Weise mit einer Rate v nach A mutiert. In diesem Fall stellt sich langfristig ein Gleichgewicht zwischen Hin- und Rückmutation und genetischer Drift ein. Die Frequenz von A hat dann eine stationäre, d. h. zeitunabhängige Verteilung. Der Erwartungswert dieser Verteilung beträgt $v/(v+v)$, d. h. der Anteil der Mutationsrate von a nach A an der Summe der Mutationsraten. Wegen der zufälligen Natur des Reproduktionsmechanismus ist die Frequenz von Allel A jedoch nicht konstant, sondern schwankt von Generation zu Generation mit einer Varianz von ungefähr $v^2/4N \times (v+v)^3$ um ihren Erwartungswert [3]. Diese Schwankung ist umso geringer, je größer N ist. Der durch Hin- und Rückmutation entstehende Polymorphismus ist also ein Dauerzustand und kein Übergangszustand wie im Fall genetischer Drift allein.

Selektion

Da an anderer Stelle in dieser Ausgabe ausführlich auf das Thema Selektion eingegangen wird, soll hier nur der Spezialfall der gerichteten Selektion eines diallelischen Gens im Wright-Fisher-Modell betrachtet werden. Dabei nehmen wir an, dass die relative Fitness der Genotypen

AA, Aa und aa jeweils $1+s$, $1+s/2$ und 1 beträgt. Der Selektionskoeffizient s sei klein und positiv, d. h. das Allel A soll einen additiven Selektionsvorteil besitzen. Wenn Mutationen von A nach a bzw. von a nach A stattfinden, wird sich in der Population ein Mutations-Selektions-Gleichgewicht einstellen. Ohne Mutation hingegen wird wegen der genetischen Drift eines der beiden Allele fixiert. Bei einer Anfangsfrequenz von π für das Allel A beträgt die Wahrscheinlichkeit, dass A fixiert wird, approximativ

$$P(\pi) = \frac{1 - e^{-2Ns\pi}}{1 - e^{-2Ns}}$$

Tab. 1 gibt für einige Werte von n , s und π die Fixierungswahrscheinlichkeiten p des Allels A an. Man erkennt, dass Selektion für große Werte von $N \times s$ wie $10^4 \times 10^{-3} = 10$ die genetische Drift dominiert. Ein Allel mit einer Anfangsfrequenz von 0,5, die gleichzeitig der Fixierungswahrscheinlichkeit für $s=0$ entspricht, wird mit an Sicherheit grenzender Wahrscheinlichkeit ($p > 0,9999$) fixiert, aber auch ein vergleichsweise seltenes Allel kann in einer großen Population seinen Selektionsvorteil ausschöpfen. Selbst die Fixierungswahrscheinlichkeit einer Neumutation, die ja nur eine Anfangsfrequenz von $1/2N$ hat, ist bei $N=10^4$ und $N \times s=10$ etwa 20-mal größer als ohne Selektion. Noch eindrucksvoller ist der Effekt der Selektion, wenn sich ein Allel zunächst durch genetische Drift anreichert und dann einen schwachen selektiven Vorteil erhält (z. B. durch Veränderung der Umweltbedingungen). Für $\pi=0,05$, $s=10^{-3}$ und $N=10^4$ erhält man eine Fixierungswahrscheinlichkeit von 0,6321, verglichen mit 0,05 ohne Selektion (da $\pi=0,05$). Für kleine Werte von $N \times s$ (z. B. $10^3 \times 10^{-4} = 0,1$) dominiert die genetische Drift die Selektion, und die Fixierungswahrscheinlichkeiten mit und ohne Selektion sind ungefähr gleich groß.

Unendlich viele Allele

In diesem Abschnitt betrachten wir ein grundlegend anderes Mutationsmodell als bisher. Im so genannten „infinite alleles model“ (IAM) wird angenommen, dass jede Mutation ein neues Allel erzeugt, d. h. es gibt keine Rückmutation

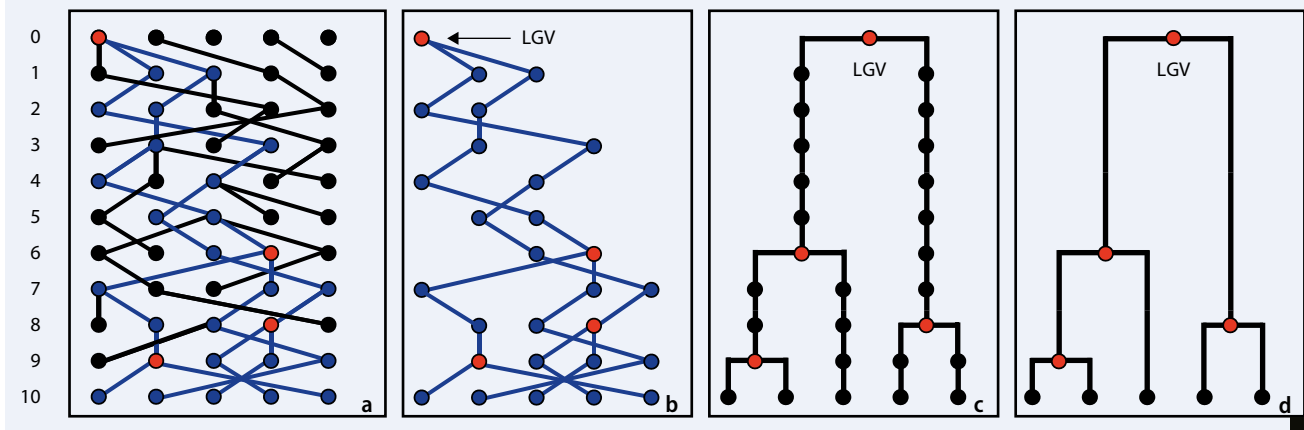


Abb. 3 ▲ Approximation des Wright-Fisher-Prozesses durch den Koaleszenzprozess, **a** Realisierung des Wright-Fisher-Prozesses, **b** zu **a** gehöriger Stammbaum der 5 Gene der rezentesten Generation 10, **c** vereinfachte Darstellung des Stammbaums aus **b**, **d** zu **c** gehöriger Koaleszenzprozess, rote Kreise Koaleszenzereignisse, LGV letzter gemeinsamer Vorfahr der 5 Gene der rezentesten Generation 10

von einem existierenden Allel zu einem anderen bereits existierenden Allel [10]. Man geht also von unendlich vielen verschiedenen Alleltypen aus. Dieses Modell lässt sich dadurch motivieren, dass bei einer durchschnittlichen Genlänge von etwa 1500 Nukleotiden die Wahrscheinlichkeit dafür, dass ein (rekurrenter) Einzelbasenaustausch zu einem bereits existierenden Allel des Gens führt, sehr gering ist. Aus diesem Grund wird das IAM häufig bei der Untersuchung von Allozymen oder Proteinpolymorphismen bzw. zur Analyse umfangreicher Haplotypen angewandt.

Unter Annahme des IAM wird im Wright-Fisher-Modell jedes neu entstandene Allel mit hoher Wahrscheinlichkeit durch genetische Drift wieder aus der Population entfernt. Allerdings drängen immer wieder neue Mutationen nach, sodass die Häufigkeitsverteilung der vorhandenen Alleltypen mit der Zeit einem Gleichgewichtszustand zustrebt. Das bedeutet, dass die Wahrscheinlichkeit für das k -malige Auftreten eines existierenden Alleltyps ($k=1, \dots, N$) nach hinreichend langer Zeit nahezu konstant ist. Die entsprechenden Wahrscheinlichkeiten werden durch die berühmte „Ewens Sampling Formula“ gegeben [2]. Insbesondere ist die erwartete Anzahl verschiedener Alleltypen in einer Population eine feste Größe. Da in der Praxis statt der gesamten Population meistens nur eine Stichprobe aus der Population mit ihrer genetischen Information zur Verfügung steht, formulieren wir das eben genannte Resultat für ei-

ne Stichprobe der Größe n aus einer Population der Größe $2N$. Sei μ die Mutationsrate pro Generation und K die Anzahl der existierenden verschiedenen Alleltypen in der Stichprobe. Bezeichnet $\theta=4N\mu$ die skalierte Mutationsrate, ergibt sich

$$E(K) = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \dots + \frac{\theta}{\theta+n-1}$$

Eine schöne Eigenschaft des IAM besteht darin, dass sich mit obiger Formel die skalierte Mutationsrate θ leicht aus einer Stichprobe schätzen lässt. Man setzt dazu einfach die beobachtete Anzahl der Alleltypen in der Stichprobe gleich $E(K)$ und löst die Gleichung nach θ auf. Dies lässt sich jedoch nicht „zu Fuß“ sondern nur mit Hilfe eines Computers bewerkstelligen.

Unendlich viele Mutationsstellen

Das „infinite sites model“ (ISM) ist ein Spezialfall des IAM aus dem vorherigen Abschnitt [9]. Beim ISM werden Gene als Sequenzen von Nukleotiden betrachtet. Es wird angenommen, dass jede Mutation eine andere Position eines Gens trifft (wodurch insbesondere auch immer ein neues Allel entsteht). Man geht also von unendlich vielen möglichen Mutationspositionen aus. Im Gegensatz zum IAM kann aber nicht nur die Anzahl der verschiedenen Alleltypen, sondern auch die Anzahl der diese Allele unterscheidenden Positionen, der so genannten „segregating sites“, berücksichtigt werden. Das ISM ist also z. B. dafür geeignet, die Evolution von

DNA-Sequenzen im Hinblick auf das Auftreten von SNP („single nucleotide polymorphism“) zu modellieren.

Eine Größe von besonderem populationsgenetischem Interesse ist die Anzahl der „segregating sites“ im ISM. Es bezeichne wiederum μ die Mutationsrate und $\theta=4N\mu$ die skalierte Mutationsrate. Dann ergibt sich die erwartete Anzahl S_n der „segregating sites“ in einer Stichprobe vom Umfang n aus folgender einfachen Formel:

$$E(S_n) = \theta \sum_{j=1}^{n-1} \frac{1}{j}$$

Mit Hilfe dieser Gleichung ist es ebenfalls möglich, θ aus einer Stichprobe zu schätzen, wenn man die Anzahl der „segregating sites“ im untersuchten Gen kennt. Vergleicht man die Schätzer von θ aus dem IAM und dem ISM, weist der Schätzer aus dem IAM immer eine größere Varianz auf als der aus dem ISM. Das liegt daran, dass im ISM mehr Information verwendet wird als im IAM, wodurch eine größere Präzision des Schätzers erreicht wird. Falls es die Datenlage zulässt, sollte also das ISM dem IAM in der praktischen Anwendung vorgezogen werden.

Kingman-Koaleszenzmodell

Grundlage vieler populationsgenetischer Überlegungen ist die Genealogie der untersuchten Gene. Ein mögliches Modell für diese ist das im Vorangegangenen vorgestellte Wright-Fisher-Modell. Ein anderes zentrales Modell, das zudem ei-

nen besonderen inhaltlichen und mathematischen Reiz hat, ist das 1982 von John F.C. Kingman eingeführte Koaleszenzmodell [11]. Betrachtet werden dabei wiederum eine diploide Population der Größe N und eine Stichprobe von n Genen aus dieser Population, wobei N sehr viel größer als n angenommen wird. Der Koaleszenzprozess startet zum aktuellen Zeitpunkt mit n Genen und verläuft, anders als der Wright-Fisher-Prozess, rückwärts in der Zeit (■ **Abb. 3d**). Der nächste Zeitpunkt von Interesse ist der, zu dem letztmalig zwei der n Gene einen gemeinsamen Vorfahren hatten. Man spricht von einem „Verschmelzen“ der beiden Gene zu diesem Zeitpunkt oder von einem „Koaleszenzereignis“. Grafisch drückt sich dieser Verschmelzungsvorgang dadurch aus, dass die Abstammungslinien der beiden Gene ab dem Koaleszenzzeitpunkt zusammenfallen. Davor sind nur noch $n-1$ Linien vorhanden, die $n-1$ Genen entsprechen. Für diese Gene wiederholt sich bei weiter rückwärts laufender Zeit der geschilderte Mechanismus, d. h. zwei der $n-1$ Gene haben wiederum letztmalig einen gemeinsamen Vorfahren gehabt, und zu diesem Zeitpunkt verschmelzen ihre beiden Linien. Der Prozess setzt sich fort, bis nur noch ein einziges Gen übrig ist, das dem letzten (oder jüngsten) gemeinsamen Vorfahren aller Gene der Ausgangsgeneration entspricht.

Der Koaleszenzprozess wird durch zwei stochastische Vorgänge charakterisiert, nämlich die Auswahl der jeweiligen Koaleszenzpartner und die Zeit zwischen den Koaleszenzereignissen. Die Auswahl der Koaleszenzpartner wird in der Regel als zufällig vorausgesetzt, d. h. jedes vorhandene Gen hat die gleiche Wahrscheinlichkeit, am nächsten Koaleszenzereignis teilzunehmen. Der Koaleszenzprozess geht also von einer panmiktischen Population mit „random mating“ und ohne Selektion aus. Zum zeitlichen Verlauf des Prozesses ist zunächst zu bemerken, dass die Zeit im Koaleszenzprozess stetig ist und nicht, wie im Wright-Fisher-Prozess, in diskreten Generationenschritten gemessen wird. Um für verschiedene Populationsgrößen N anwendbar zu sein, wird die Zeit im Koaleszenzmodell außerdem normiert (das bedeutet, dass die Prozesszeit erst mit $2N$ multipliziert wer-

den muss, um die Zeit in Generationen zu erhalten). Für n betrachtete Gene folgt die Zeit bis zum nächsten Koaleszenzereignis einer Exponentialverteilung mit Parameter $n \times (n-1)/2$, d. h. die mittlere Koaleszenzzeit beträgt $2/[n \times (n-1)]$. Das bedeutet wiederum, dass die Zeit bis zum Verschmelzen zweier Gene immer länger wird, je weniger Gene vorhanden sind. Bei 2 Genen beträgt die mittlere Zeit bis zum gemeinsamen Vorfahren 1, also $2N$ Generationen. Liegt eine Stichprobe von n Genen vor, beträgt die erwartete Zeit bis zum letzten gemeinsamen Vorfahren dieser Gene, d. h. bis zum letzten Verschmelzen der letzten beiden Linie, $2-2/n$, also ungefähr 2 bei nicht zu kleinem n . Dies bedeutet, dass auch bei einer sehr großen Stichprobe die Zeit bis zum Erreichen des letzten gemeinsamen Vorfahrens der Gene gerade einmal doppelt so lang ist wie bei zwei Genen. Es lässt sich zudem mathematisch beweisen, dass die Varianz dieser Zeitspanne im Wesentlichen von der Varianz der Koaleszenzzeit der letzten beiden Gene bestimmt wird.

Der Koaleszenzprozess ist der Grenzprozess des Wright-Fisher-Prozesses für große Populationsgrößen (■ **Abb. 3**). Ähnliches gilt für weitaus kompliziertere Modelle des Reproduktionsprozesses wie das Cannings- und das Moran-Modell. Neben der Ermöglichung qualitativer und quantitativer Aussagen über die zeitliche Entwicklung von Populationen hat der Koaleszenzprozess den großen Vorteil, dass er sehr einfach simuliert werden kann. Es ist nicht nötig, wie beim Wright-Fisher-Prozess jede einzelne Generation vollständig zu simulieren; die Simulation muss nur die Koaleszenzereignisse erfassen. Die hierfür notwendigen Rechenschritte können auch vorwärts in der Zeit ausgeführt werden, um damit die zukünftige Evolution von Populationen zu untersuchen. Einfacher und gebräuchlicher ist es aber, die Simulationen wie oben geschildert rückwärts in der Zeit durchzuführen. Dabei wird von Daten aus der Jetztzeit ausgegangen und versucht, Schlüsse über Vorgänge in der Vergangenheit zu ziehen. In jüngster Zeit erfreuen sich solche Simulationen in der Populationsgenetik großer Beliebtheit, da damit realistischere Parameterschätzungen und statistische Tests durchgeführt wer-

den können. Meist wird in den Simulationen auch angenommen, dass der Mutationsprozess unabhängig von der Genealogie abläuft, d. h. dass Mutationen unabhängig von der Struktur des Koaleszenzbaums simuliert und dann auf diesen projiziert werden können.

Ausblick

Das Wright-Fisher-Modell basiert auf Voraussetzungen, die in der Realität normalerweise nicht erfüllt sind. Deshalb gibt es in der mathematischen Populationsgenetik inzwischen zahlreiche Weiterentwicklungen. Hierzu gehört z. B. das Moran-Modell [12], in dem die Annahme diskreter, nicht überlappender Generationen aufgegeben wird. In ihm können Gene zu einem beliebigen Zeitpunkt „sterben“ und werden sofort durch ein neues, evtl. mutiertes Gen ersetzt. Eine wichtige Verallgemeinerung des Wright-Fisher-Modells ist das Cannings-Modell [1], bei dem sich Gene unter Berücksichtigung der fest vorgegebenen Populationsgröße N beliebig „fortpflanzen“ können. Voraussetzung ist dabei nur, dass die Verteilung der Nachkommen von n Genen allein von dieser Zahl n abhängt, nicht aber von der speziellen Auswahl der Gene. Insbesondere werden auch stochastische Abhängigkeiten zwischen Genen zugelassen. Viele Ergebnisse des Wright-Fisher-Modells lassen sich auf das Cannings-Modell übertragen, wobei die Varianz der Zahl der Nachkommen eines Gens ein zusätzlicher Parameter des Modells ist. Andere Modelle berücksichtigen Art und Änderung der räumlichen Struktur einer Population, wobei der Begriff des so genannten „gene flow“ eine maßgebliche Rolle spielt.

Da das Koaleszenzmodell sehr häufig in der Praxis angewandt wird, wurden für dieses auch eine Reihe von Verallgemeinerungen vorgeschlagen, die das Modell besser an die Realität anpassen sollen [13]. Mit ihrer Zuhilfenahme lässt sich eine variable Bevölkerungsgröße, z. B. in der Form exponentiellen Wachstums, ebenso in das Koaleszenzmodell einbeziehen wie der Geschlechtsdimorphismus oder das Auftreten von Selektion.

Theoretische populationsgenetische Modelle haben unser Verständnis der

Evolution von Populationen unter der Berücksichtigung von Mutation und Selektion erheblich erweitert. Zudem haben sie uns die Möglichkeit eröffnet, aus den vielfältig vorhandenen genetischen Daten Schlüsse über die ihnen zugrunde liegenden Mechanismen und Parameter (wie Mutationsraten) zu ziehen. Bei alledem darf allerdings nicht vergessen werden, dass die Realität evolutionärer Vorgänge – zumal beim Menschen – extrem kompliziert ist. Jedes Modell kann daher nur als grobe Approximation der tatsächlichen Gegebenheiten verstanden werden. Für ein und denselben Datensatz eignen sich oftmals mehrere Modelle, und dementsprechend führt deren Anwendung auch zu unterschiedlichen Schätzungen von Populationsparametern und zu unterschiedlichen statistischen Schlussfolgerungen. Diese Modellunsicherheit spiegelt sich nicht in Konfidenzintervallen oder p-Werten wider, denn es wird davon ausgegangen, dass das jeweils zugrunde gelegte Modell korrekt ist. Bei der Interpretation und der Bewertung von Ergebnissen, die mittels eines populationsgenetischen Modells gewonnen wurden, sind die Probleme der Modellwahl und des nur approximativen Charakters jedes Modells immer zu berücksichtigen.

Korrespondenzadresse

Dr. A. Caliebe

Institut für Medizinische Informatik
und Statistik, Haus 31,
Universitätsklinikum Schleswig-Holstein,
Campus Kiel,
Arnold-Heller-Straße 3, 24105 Kiel
caliebe@medinfo.uni-kiel.de

Interessenkonflikt. Der korrespondierende Autor gibt an, dass kein Interessenkonflikt besteht.

Literatur

1. Cannings C (1974) The latent roots of certain Markov chains arising in genetics: a new approach 1. Haploid models. *Adv Appl Prob* 6: 260–290
2. Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theor Popul Biol* 3: 87–112
3. Ewens WJ (2004) *Mathematical population genetics*, 2nd edn. Springer, Berlin Heidelberg New York
4. Fisher RA (1922) On the dominance ratio. *Proc R Soc Edinb* 42: 321–341
5. Fisher RA (1930) *The genetical theory of natural selection*. Clarendon Press, Oxford
6. Haldane JBS (1924) A mathematical theory of natural and artificial selections, Part II. *Proc Camb Phil Soc Biol Sci* 1: 158–163
7. Haldane JBS (1937) The effect of variation on fitness. *Am Nat* 71: 337–349
8. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217: 624–626
9. Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. *Genetics* 61: 893
10. Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49: 725–738
11. Kingman JFC (1982) The coalescent. *Stochastic Proc Appl* 13: 235–248
12. Moran PAP (1958) Random processes in genetics. *Proc Camb Phil Soc* 54: 60–71
13. Nordborg M (2007) Coalescent theory. In: Balding DJ, Bishop M, Cannings C (eds) *Statistical genetics*, 3rd edn. Wiley, Chichester, pp 843–877
14. Ohta T, Kimura M (1969) Linkage disequilibrium due to random genetic drift. *Genet Res* 13: 47–55
15. Wright S (1931) Evolution in Mendelian populations. *Genetics* 16: 97–159



- Kongressnews
 - Spannendes aus der Welt der Medizin
 - Interviews
- Jeden Monat neu!**

Jetzt kostenlos downloaden unter

www.springer.de/podcast